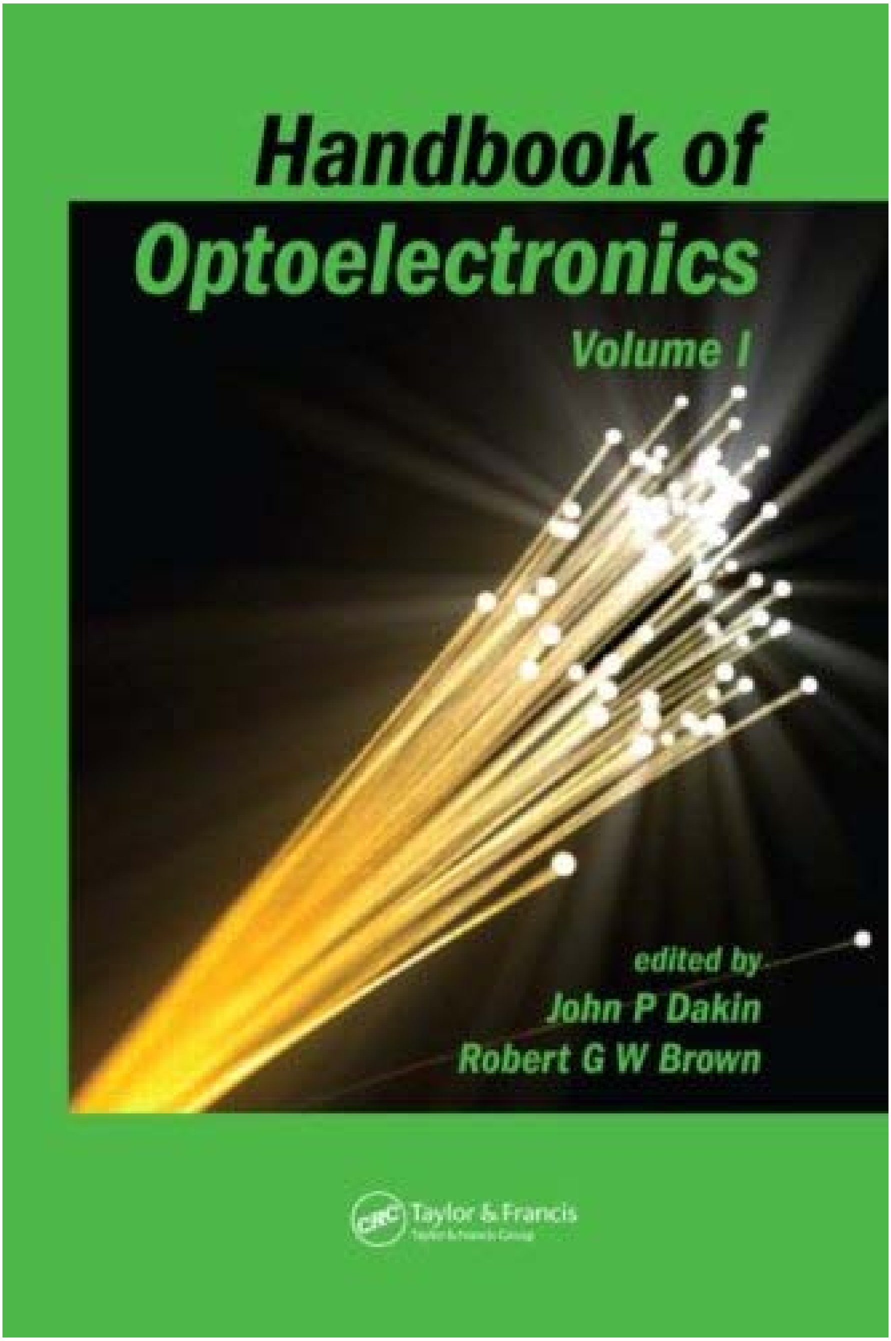
Handbook of Optoelectronics


Handbook of
Optoelectronics
Volume I
edited by
John P Dakin
Robert G W Brown
CRC Taylor & Francis
Taylor & Francis Group

VISIT…

LANZAROTE
Caliente.COM

Handbook of Optoelectronics

Volume I

Handbook of Optoelectronics

Volume I

edited by

John P Dakin
University of Southhampton, UK

Robert G W Brown
University of Nottingham, UK

Taylor & Francis
Taylor & Francis Group
New York London

Taylor & Francis is an imprint of the
Taylor & Francis Group, an informa business

Published in 2006 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2006 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-7503-0646-7 (Hardcover)
International Standard Book Number-13: 978-0-7503-0646-1 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Editorial Board

Editors-in-Chief
John P Dakin and Robert G W Brown

Section Editors

Roel Baets
Ghent University–IMEC
Ghent, Belgium

Jean-Luc Beylat
Alcatel
Villarceaux, France

Edward Browell
NASA Langley Research Center
Hampton, Virginia

Robert G W Brown
Department of Electronic Engineering
University of Nottingham
Nottingham, United Kingdom

John P Dakin
Optoelectronics Research Centre
University of Southampton
Highfield, United Kingdom

Michel Digonnet
Edward L. Gintzon Laboratory
Stanford University
Stanford, California

Galina Khitrova
College of Optical Sciences
University of Arizona
Tucson, Arizona

Peter Raynes
Department of Engineering
University of Oxford
Oxford, United Kingdom

Alan Rogers
Department of Electronic Engineering
University of Surrey
Guildford, United Kingdom

Tatsuo Uchida
Department of Electronics Engineering
Tohoku University
Sendai, Japan

List of contributors

Takao Ando
Shizuoka University
Shizuoka, Japan

Nicholas Baynes
CDT Limited
Cambridge, United Kingdom

Zbigniew Bielecki
Military University of Technology
Warsaw, Poland

Anders Bjarklev
Bjarklev Consult APS
Roskilde, Denmark

Nikolaus Boos
EADS Eurocopter SAS
Marignane, France

Jens Buus
Gayton Photonics
Gayton, United Kingdom

Chien-Jen Chen
Onetta Inc.
San Jose, California

Dominique Chiaroni
Alcatel CIT
Marcoussis, France

Krzysztof Chrzanowski
Military University of Technology
Warsaw, Poland

David Coates
CRL
Hayes, United Kingdom

Nadir Dagli
University of California
Santa Barbara, California

John Dakin
ORC, Southampton University
Southampton, United Kingdom

Xavier Daxhelet
Ecole Polytechnique de Montreal
Montreal, Quebec, Canada

Michel Digonnet
Stanford University
Stanford, California

Uzi Efron
Ben-Gurion University
Beer-Sheva, Israel

Günter Gauglitz
Institut für Physikalische und
Theoretische Chemie
Tübingen, Germany

Ron Gibbs
Gibbs Associates
Dunstable, United Kingdom

Martin Grell
University of Sheffield
Sheffield, United Kingdom

Nick Holliman
University of Durham,
Durham, United Kingdom

Kazuo Hotate
University of Tokya
Tokyo, Japan

Michel Joindot
France Telecom
Lannion, France

George K Knopf
The University of Western Ontario
London, Ontario, Canada

Ton Koonen
Eindhoven University of Technology
Eindhoven, The Netherlands

Hidehiro Kume
Hamamatsu Photonics KK
Shizuoka, Japan

Suzanne Lacroix
Ecole Polytechnique de Montreal
Montreal, Quebec, Canada

Jesper Lægsgaard
University of Southampton
Southampton, United Kingdom

John N Lee
Naval Research Laboratory
Washington, District of Columbia

Christian Lerminiaux
Corning SA – CERF
Avon, France

Robert A Lieberman
Intelligent Optical Systems Inc.
Torrance, California

John Love
Australian National University
Canberra, Australia

Makoto Maeda
Home Network Company, SONY
Kanagawa, Japan

Michael A Marcus
Eastman Kodak Company
Rochester, New York

Tom Markvart
University of Southampton
Southampton, United Kingdom

Tanya M Monro
University of Southampton
Southampton, United Kingdom

Johan Nilsson
University of Southampton
Southampton, United Kingdom

Yoshi Ohno
National Institute of Standards
and Technology
Gaithersburg, Maryland

Susanna Orlic
Technical University Berlin
Berlin, Germany

Antoni Rogalski
Military University of Technology
Warsaw, Poland

Alan Rogers
University of Surrey
Guildford, United Kingdom

Neil Ross
University of Southampton
Southampton, United Kingdom

Tsutae Shinoda
Fujitsu Laboratories Ltd
Akashi, Japan

Hilary G Sillitto
Edinburgh, United Kingdom

Anthony E Smart
Scattering Solutions, LLC
Costa Mesa, California

Brian Smith
Pips Technology
Hampshire, United Kingdom

Euan Smith
CDT Limited
Cambridge, United Kingdom

Peter G R Smith
University of Southampton
Southampton, United Kingdom

Günter Steinmeyer
Max-Born-Institute for Nonlinear Optics and Short Pulse Spectroscopy
Berlin, Germany

Klaus Streubel
OSRAM Opto Semiconductors
Regensberg, Germany

Masayuki Sugawara
NHK Science and Technical Research Laboratories
Tokyo, Japan

Yan Sun
Onetta Inc.
San Jose, California

Kenkicki Tanioka
NHK Science and Technical Research Laboratories
Tokyo, Japan

Heiju Uchiike
Saga Univerisity
Saga, Japan

J Michael Vaughan
Research Consultant, Optoelectronics
Buckinghamshire, United Kingdom

Tuan Vo-Dinh
Oak Ridge National Laboratory
Oak Ridge, Tennessee

David O Wharmby
Technology Consultant
Ilkley, United Kingdom

William S Wong
Onetta Inc.
San Jose, California

Acknowledgments

Firstly we must thank all the many leading scientists and technologists who have contributed so generously to the chapters of this book. It is no small task to produce a comprehensive and dispassionately accurate summary, even of your own research field, and we are most grateful to all of our authors.

John Dakin would like to acknowledge all his family, close friends and colleagues at Southampton University who have been most understanding during the production of this Handbook.

Robert Brown would like to acknowledge the constant support of his wife and close family throughout the preparation of this book.

We both wish to give special thanks to the (UK) Institute of Physics Publishing (IoPP) staff who did so much during the book's development and production period. Gillian Lindsay worked hard with the two of us at the outset some years ago, and Karen Donnison took over in the middle-period and patiently cajoled the editors and authors to deliver on their promises. Lastly we thank Dr. John Navas, who took over the reins in the final stages. He carried the book forward from IoPP to Taylor and Francis in the last few months and enabled the final product to be delivered.

Robert G W Brown
John P Dakin

Introduction

Optoelectronics is a remarkably broad scientific and technological field that supports a multi-billion US-dollar per annum global industry, employing tens of thousands of scientists and engineers. The optoelectronics industry is one of the great global businesses of our time.

In this Handbook, we have aimed to produce a book that is not just a text containing theoretically-sound physics & electronics coverage, nor just a practical engineering handbook, but a text designed to be strong in both these areas. We believe that, with the combined assistance of many world experts, we have succeeded in achieving this very difficult aim. The structure and contents of this Handbook have proved fascinating to assemble, using this input from so many leading practitioners of the science, technology and art of optoelectronics.

Today's optical telecommunications, display and illumination technologies rely heavily on optoelectronic components: laser diodes, light emitting diodes, liquid crystal and plasma screen displays etc. In today's world it is virtually impossible to find a piece of electrical equipment that does not employ optoelectronic devices as a basic necessity – from CD and DVD players to televisions, from automobiles and aircraft to medical diagnostic facilities in hospitals and telephones, from satellites and space-borne missions to underwater exploration systems – the list is almost endless. Optoelectronics is in virtually every home and business office in the developed modern world, in telephones, fax machines, photocopiers, computers and lighting.

'Optoelectronics' is not precisely defined in the literature. In this Handbook we have covered not only optoelectronics as a subject concerning devices and systems that are essentially electronic in nature, yet involve light (such as the laser diode), but we have also covered closely related areas of electro-optics, involving devices that are essentially optical in nature but involve electronics (such as crystal light-modulators).

To provide firm foundations, this Handbook opens with a section covering 'Basic Concepts'. The 'Introduction' is followed immediately by a chapter concerning 'Materials', for it is through the development and application of new materials and their special properties that the whole business of optoelectronic science and technology now advances. Many optoelectronic systems still rely on conventional light sources rather than semiconductor sources, so we cover these in the third chapter, leaving semiconductor matters to a later section. The detection of light is fundamental to many optoelectronic systems, as are optical waveguides, amplifiers and lasers, so we cover these in the remaining chapters of the Basic Concepts section.

The 'Advanced Concepts' section focuses on three areas that will be useful to some of our intended audience, both now, in advanced optics and photometry – and now and increasingly in the future concerning non-linear and short-pulse effects.

'Optoelectronics Devices and Techniques' is a core foundation section for this Handbook, as today's optoelectronics business relies heavily on such knowledge. We have attempted to cover all the main areas of semiconductor optoelectronics devices and materials in the eleven chapters in this section, from light emitting diodes and lasers of great variety to fibers, modulators and amplifiers. Ultra-fast and integrated devices are increasingly important, as are organic electroluminescent devices and photonic bandgap and crystal fibers. Artificially engineered materials provide a rich source of possibility for next generation optoelectronic devices.

At this point the Handbook ‘changes gear’ – and we move from the wealth of devices now available to us – to how they are used in some of the most important optoelectronic systems available today. We start with a section covering ‘Communication’, for this is how the developed world talks and communicates by internet and email today – we are all now heavily dependent on optoelectronics. Central to such optoelectronic systems are transmission, network architecture, switching and multiplex architectures – the focus of our chapters here. In Communication we already have a multi-tens-of-billions-of-dollars-per-annum industry today.

‘Imaging and displays’ is the other industry measured in the tens of billions of dollars per annum range at the present time. We deal here with most if not all of the range of optoelectronic techniques used today from cameras, vacuum and plasma displays to liquid crystal displays and light modulators, from electroluminescent displays and exciting new 3-dimensional display technologies just entering the market place in mobile telephone and laptop computer displays – to the very different application area of scanning and printing.

‘Sensing and Data Processing’ is a growing area of optoelectronics that is becoming increasingly important – from non-invasive patient measurements in hospitals to remote sensing in nuclear power stations and aircraft. At the heart of many of today’s sensing capabilities is the business of optical fiber sensing, so we begin this section of the Handbook there, before delving into remote optical sensing and military systems (at an un-classified level – for here-in lies a problem for this Handbook – that much of the current development and capability in military optoelectronics is classified and un-publishable because of it’s strategic and operational importance). Optical information storage and recovery is already a huge global industry supporting the computer and media industries in particular; optical information processing shows promise but has yet to break into major global utilization. We cover all of these aspects in our chapters here.

‘Industrial Medical and Commercial Applications’ of optoelectronics abound and we cannot possibly do justice to all the myriad inventive schemes and capabilities that have been developed to date. However, we have tried hard to give a broad overview within major classification areas, to give you a flavor of the sheer potential of optoelectronics for application to almost everything that can be measured. We start with the foundation areas of spectroscopy – and increasingly important surveillance, safety and security possibilities. Actuation and control – the link from optoelectronics to mechanical systems is now pervading nearly all modern machines: cars, aircraft, ships, industrial production etc – a very long list is possible here. Solar power is and will continue to be of increasing importance – with potential for urgently needed breakthroughs in photon to electron conversion efficiency. Medical applications of optoelectronics are increasing all the time, with new learned journals and magazines regularly being started in this field.

Finally we come to the art of practical optoelectronic systems – how do you put optoelectronic devices together into reliable and useful systems, and what are the ‘black art’ experiences learned through painful experience and failure? This is what other optoelectronic books never tell you – and we are fortunate to have a chapter that addresses many of the questions we should be thinking about as we design and build systems – but often forget or neglect at our peril.

In years to come, optoelectronics will develop in many new directions. Some of the more likely directions to emerge by 2010 will include optical packet switching, quantum cryptographic communications, three-dimensional and large-area thin-film displays, high-efficiency solar-power generation, widespread bio-medical and bio-photonic disease analyses and treatments and optoelectronic purification processes. Many new devices will be based on quantum dots, photonic crystals and nano-optoelectronic components. A future edition of this Handbook is likely to report on these rapidly changing fields currently pursued in basic research laboratories.

We are confident you will enjoy using this Handbook of Optoelectronics, derive fascination and pleasure in this richly rewarding scientific and technological field, and apply your knowledge in either your research or your business.

Robert G W Brown
John P Dakin

Table of Contents

BASIC CONCEPTS *Alan Rogers*

A1.1 An introduction to optoelectronics *Alan Rogers* 1
A1.2 Optical materials *Neil Ross* 21
A1.3 Incandescent, discharge and arc lamp sources *David O Wharmby* 45
A1.4 Detection of optical radiation *Antoni Rogalski and Zbigniew Bielecki* 73
A1.5 Propagation along optical fibres and waveguides *John Love* 119
A1.6 Introduction to lasers and optical amplifiers *William S Wong, Chien-Jen Chen and Yan Sun* 179

ADVANCED CONCEPTS *Alan Rogers and Galina Khitrova*

A2.1 Advanced optics *Alan Rogers* 205
A2.2 Basic concepts in photometry, radiometry and colorimetry *Yoshi Ohno* 287
A2.3 Nonlinear and short pulse effects *Günter Steinmeyer* 307

OPTOELECTRONIC DEVICES AND TECHNIQUES *John P Dakin, Roel Bates and Michel Digonnet*

B1.1 Visible light-emitting diodes *Klaus Streubel* 329
B1.2 Semiconductor lasers *Jens Buus* 385
B2 Optical detectors and receivers *Hidehiro Kume* 413
B3 Optical fibre devices *Suzanne Lacroix and Xavier Daxhelet* 457
B4 Optical modulators *Nadir Dagli* 489
B5 Optical amplifiers *Johan Nilsson, Jesper Lægsgaard and Anders Bjarklev* 533
B6 Ultrafast optoelectronics *Günter Steinmeyer* 565
B7 Integrated optics *Nikolaus Boos and Christian Lerminiaux* 587
B8 Infrared devices and techniques *Antoni Rogalski and Krzysztof Chrzanowski* 653
B9 Organic light emitting devices *Martin Grell* 693
B10 Microstructured optical fibres *Tanya M Monro, Anders Bjarklev and Jesper Lægsgaard* 719
B11 Engineered optical materials *Peter G R Smith* 745

Handbook of Optoelectronics

Volume II

Handbook of Optoelectronics

Volume II

edited by

John P Dakin
University of Southhampton, UK

Robert G W Brown
University of Nottingham, UK

Taylor & Francis
Taylor & Francis Group
New York London

Taylor & Francis is an imprint of the
Taylor & Francis Group, an informa business

Published in 2006 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

CRC Press is an imprint of Taylor & Francis Group

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-7503-0646-7 (Hardcover)
International Standard Book Number-13: 978-0-7503-0646-1 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Editorial Board

Editors-in-Chief
John P Dakin and Robert G W Brown

Section Editors

Roel Baets
Ghent University–IMEC
Ghent, Belgium

Jean-Luc Beylat
Alcatel
Villarceaux, France

Edward Browell
NASA Langley Research Center
Hampton, Virginia

Robert G W Brown
Department of Electronic Engineering
University of Nottingham
Nottingham, United Kingdom

John P Dakin
Optoelectronics Research Centre
University of Southampton
Highfield, United Kingdom

Michel Digonnet
Edward L. Gintzon Laboratory
Stanford University
Stanford, California

Galina Khitrova
College of Optical Sciences
University of Arizona
Tucson, Arizona

Peter Raynes
Department of Engineering
University of Oxford
Oxford, United Kingdom

Alan Rogers
Department of Electronic Engineering
University of Surrey
Guildford, United Kingdom

Tatsuo Uchida
Department of Electronics Engineering
Tohoku University
Sendai, Japan

List of contributors

Takao Ando
Shizuoka University
Shizuoka, Japan

Nicholas Baynes
CDT Limited
Cambridge, United Kingdom

Zbigniew Bielecki
Military University of Technology
Warsaw, Poland

Anders Bjarklev
Bjarklev Consult APS
Roskilde, Denmark

Nikolaus Boos
EADS Eurocopter SAS
Marignane, France

Jens Buus
Gayton Photonics
Gayton, United Kingdom

Chien-Jen Chen
Onetta Inc.
San Jose, California

Dominique Chiaroni
Alcatel CIT
Marcoussis, France

Krzysztof Chrzanowski
Military University of Technology
Warsaw, Poland

David Coates
CRL
Hayes, United Kingdom

Nadir Dagli
University of California
Santa Barbara, California

John Dakin
ORC, Southampton University
Southampton, United Kingdom

Xavier Daxhelet
Ecole Polytechnique de Montreal
Montreal, Quebec, Canada

Michel Digonnet
Stanford University
Stanford, California

Uzi Efron
Ben-Gurion University
Beer-Sheva, Israel

Günter Gauglitz
Institut für Physikalische und
Theoretische Chemie
Tübingen, Germany

Ron Gibbs
Gibbs Associates
Dunstable, United Kingdom

Martin Grell
University of Sheffield
Sheffield, United Kingdom

Nick Holliman
University of Durham,
Durham, United Kingdom

Kazuo Hotate
University of Tokya
Tokyo, Japan

Michel Joindot
France Telecom
Lannion, France

George K Knopf
The University of Western Ontario
London, Ontario, Canada

Ton Koonen
Eindhoven University of Technology
Eindhoven, The Netherlands

Hidehiro Kume
Hamamatsu Photonics KK
Shizuoka, Japan

Suzanne Lacroix
Ecole Polytechnique de Montreal
Montreal, Quebec, Canada

Jesper Lægsgaard
University of Southampton
Southampton, United Kingdom

John N Lee
Naval Research Laboratory
Washington, District of Columbia

Christian Lerminiaux
Corning SA – CERF
Avon, France

Robert A Lieberman
Intelligent Optical Systems Inc.
Torrance, California

John Love
Australian National University
Canberra, Australia

Makoto Maeda
Home Network Company, SONY
Kanagawa, Japan

Michael A Marcus
Eastman Kodak Company
Rochester, New York

Tom Markvart
University of Southampton
Southampton, United Kingdom

Tanya M Monro
University of Southampton
Southampton, United Kingdom

Johan Nilsson
University of Southampton
Southampton, United Kingdom

Yoshi Ohno
National Institute of Standards
and Technology
Gaithersburg, Maryland

Susanna Orlic
Technical University Berlin
Berlin, Germany

Antoni Rogalski
Military University of Technology
Warsaw, Poland

Alan Rogers
University of Surrey
Guildford, United Kingdom

Neil Ross
University of Southampton
Southampton, United Kingdom

Tsutae Shinoda
Fujitsu Laboratories Ltd
Akashi, Japan

Hilary G Sillitto
Edinburgh, United Kingdom

Anthony E Smart
Scattering Solutions, LLC
Costa Mesa, California

Brian Smith
Pips Technology
Hampshire, United Kingdom

Euan Smith
CDT Limited
Cambridge, United Kingdom

Peter G R Smith
University of Southampton
Southampton, United Kingdom

Günter Steinmeyer
Max-Born-Institute for Nonlinear
Optics and Short Pulse Spectroscopy
Berlin, Germany

Klaus Streubel
OSRAM Opto Semiconductors
Regensberg, Germany

Masayuki Sugawara
NHK Science and Technical Research
Laboratories
Tokyo, Japan

Yan Sun
Onetta Inc.
San Jose, California

Kenkicki Tanioka
NHK Science and Technical Research
Laboratories
Tokyo, Japan

Heiju Uchiike
Saga Univerisity
Saga, Japan

J Michael Vaughan
Research Consultant, Optoelectronics
Buckinghamshire, United Kingdom

Tuan Vo-Dinh
Oak Ridge National Laboratory
Oak Ridge, Tennessee

David O Wharmby
Technology Consultant
Ilkley, United Kingdom

William S Wong
Onetta Inc.
San Jose, California

Acknowledgments

Firstly we must thank all the many leading scientists and technologists who have contributed so generously to the chapters of this book. It is no small task to produce a comprehensive and dispassionately accurate summary, even of your own research field, and we are most grateful to all of our authors.

John Dakin would like to acknowledge all his family, close friends and colleagues at Southampton University who have been most understanding during the production of this Handbook.

Robert Brown would like to acknowledge the constant support of his wife and close family throughout the preparation of this book.

We both wish to give special thanks to the (UK) Institute of Physics Publishing (IoPP) staff who did so much during the book's development and production period. Gillian Lindsay worked hard with the two of us at the outset some years ago, and Karen Donnison took over in the middle-period and patiently cajoled the editors and authors to deliver on their promises. Lastly we thank Dr. John Navas, who took over the reins in the final stages. He carried the book forward from IoPP to Taylor and Francis in the last few months and enabled the final product to be delivered.

Robert G W Brown
John P Dakin

Introduction

Optoelectronics is a remarkably broad scientific and technological field that supports a multi-billion US-dollar per annum global industry, employing tens of thousands of scientists and engineers. The optoelectronics industry is one of the great global businesses of our time.

In this Handbook, we have aimed to produce a book that is not just a text containing theoretically-sound physics & electronics coverage, nor just a practical engineering handbook, but a text designed to be strong in both these areas. We believe that, with the combined assistance of many world experts, we have succeeded in achieving this very difficult aim. The structure and contents of this Handbook have proved fascinating to assemble, using this input from so many leading practitioners of the science, technology and art of optoelectronics.

Today's optical telecommunications, display and illumination technologies rely heavily on optoelectronic components: laser diodes, light emitting diodes, liquid crystal and plasma screen displays etc. In today's world it is virtually impossible to find a piece of electrical equipment that does not employ optoelectronic devices as a basic necessity – from CD and DVD players to televisions, from automobiles and aircraft to medical diagnostic facilities in hospitals and telephones, from satellites and space-borne missions to underwater exploration systems – the list is almost endless. Optoelectronics is in virtually every home and business office in the developed modern world, in telephones, fax machines, photocopiers, computers and lighting.

'Optoelectronics' is not precisely defined in the literature. In this Handbook we have covered not only optoelectronics as a subject concerning devices and systems that are essentially electronic in nature, yet involve light (such as the laser diode), but we have also covered closely related areas of electro-optics, involving devices that are essentially optical in nature but involve electronics (such as crystal light-modulators).

To provide firm foundations, this Handbook opens with a section covering 'Basic Concepts'. The 'Introduction' is followed immediately by a chapter concerning 'Materials', for it is through the development and application of new materials and their special properties that the whole business of optoelectronic science and technology now advances. Many optoelectronic systems still rely on conventional light sources rather than semiconductor sources, so we cover these in the third chapter, leaving semiconductor matters to a later section. The detection of light is fundamental to many optoelectronic systems, as are optical waveguides, amplifiers and lasers, so we cover these in the remaining chapters of the Basic Concepts section.

The 'Advanced Concepts' section focuses on three areas that will be useful to some of our intended audience, both now, in advanced optics and photometry – and now and increasingly in the future concerning non-linear and short-pulse effects.

'Optoelectronics Devices and Techniques' is a core foundation section for this Handbook, as today's optoelectronics business relies heavily on such knowledge. We have attempted to cover all the main areas of semiconductor optoelectronics devices and materials in the eleven chapters in this section, from light emitting diodes and lasers of great variety to fibers, modulators and amplifiers. Ultra-fast and integrated devices are increasingly important, as are organic electroluminescent devices and photonic bandgap and crystal fibers. Artificially engineered materials provide a rich source of possibility for next generation optoelectronic devices.

At this point the Handbook ‘changes gear’ – and we move from the wealth of devices now available to us – to how they are used in some of the most important optoelectronic systems available today. We start with a section covering ‘Communication’, for this is how the developed world talks and communicates by internet and email today – we are all now heavily dependent on optoelectronics. Central to such optoelectronic systems are transmission, network architecture, switching and multiplex architectures – the focus of our chapters here. In Communication we already have a multi-tens-of-billions-of-dollars-per-annum industry today.

‘Imaging and displays’ is the other industry measured in the tens of billions of dollars per annum range at the present time. We deal here with most if not all of the range of optoelectronic techniques used today from cameras, vacuum and plasma displays to liquid crystal displays and light modulators, from electroluminescent displays and exciting new 3-dimensional display technologies just entering the market place in mobile telephone and laptop computer displays – to the very different application area of scanning and printing.

‘Sensing and Data Processing’ is a growing area of optoelectronics that is becoming increasingly important – from non-invasive patient measurements in hospitals to remote sensing in nuclear power stations and aircraft. At the heart of many of today’s sensing capabilities is the business of optical fiber sensing, so we begin this section of the Handbook there, before delving into remote optical sensing and military systems (at an un-classified level – for here-in lies a problem for this Handbook – that much of the current development and capability in military optoelectronics is classified and un-publishable because of it’s strategic and operational importance). Optical information storage and recovery is already a huge global industry supporting the computer and media industries in particular; optical information processing shows promise but has yet to break into major global utilization. We cover all of these aspects in our chapters here.

‘Industrial Medical and Commercial Applications’ of optoelectronics abound and we cannot possibly do justice to all the myriad inventive schemes and capabilities that have been developed to date. However, we have tried hard to give a broad overview within major classification areas, to give you a flavor of the sheer potential of optoelectronics for application to almost everything that can be measured. We start with the foundation areas of spectroscopy – and increasingly important surveillance, safety and security possibilities. Actuation and control – the link from optoelectronics to mechanical systems is now pervading nearly all modern machines: cars, aircraft, ships, industrial production etc – a very long list is possible here. Solar power is and will continue to be of increasing importance – with potential for urgently needed breakthroughs in photon to electron conversion efficiency. Medical applications of optoelectronics are increasing all the time, with new learned journals and magazines regularly being started in this field.

Finally we come to the art of practical optoelectronic systems – how do you put optoelectronic devices together into reliable and useful systems, and what are the ‘black art’ experiences learned through painful experience and failure? This is what other optoelectronic books never tell you – and we are fortunate to have a chapter that addresses many of the questions we should be thinking about as we design and build systems – but often forget or neglect at our peril.

In years to come, optoelectronics will develop in many new directions. Some of the more likely directions to emerge by 2010 will include optical packet switching, quantum cryptographic communications, three-dimensional and large-area thin-film displays, high-efficiency solar-power generation, widespread bio-medical and bio-photonic disease analyses and treatments and optoelectronic purification processes. Many new devices will be based on quantum dots, photonic crystals and nano-optoelectronic components. A future edition of this Handbook is likely to report on these rapidly changing fields currently pursued in basic research laboratories.

We are confident you will enjoy using this Handbook of Optoelectronics, derive fascination and pleasure in this richly rewarding scientific and technological field, and apply your knowledge in either your research or your business.

Robert G W Brown
John P Dakin

Table of Contents

COMMUNICATION *Jean-Luc Beylat*

C1.1 Optical transmission *Michel Joindot and Michel Digonnet* 765
C1.2 Optical network architectures *Ton Koonen* 797
C1.3 Optical switching and multiplexed architectures *Dominique Chiaroni* 833

IMAGING AND DISPLAYS *Peter Raynes and Tatsuo Uchida*

C2.1 Camera technology *Kenkicki Tanioka, Takao Ando and Masayuki Sugawara* 867
C2.2 Vacuum tube and plasma displays *Makoto Maeda, Tsutae Shinoda and Heiju Uchiike* 931
C2.3 Liquid crystal displays *David Coates* 957
C2.4 Technology and applications of spatial light modulators *Uzi Efron* 991
C2.5 Organic electroluminescent displays *Nicholas Baynes and Euan Smith* 1039
C2.6 Three-dimensional display systems *Nick Holliman* 1067
C2.7 Optical scanning and printing *Ron Gibbs* 1101

SENSING AND DATA PROCESSING *John P Dakin, Roel Bates and Edward Browell*

C3.1 Optical fibre sensors *John P Dakin, Kazuo Hotate, Robert A Lieberman and Michael A Marcus* 1129
C3.2 Remote optical sensing by laser *J Michael Vaughan* 1217
C3.3 Military optoelectronics *Hilary G Sillitto* 1297
C3.4 Optical information storage and recovery *Susanna Orlic* 1335
C3.5 Optical information processing *John N Lee* 1369

INDUSTRIAL, MEDICAL & COMMERCIAL APPLICATIONS *John P Dakin and Roel Bates*

C4.1 Spectroscopic analysis *Günter Gauglitz and John P Dakin* 1399
C4.2 Intelligent surveillance *Brian Smith* 1443
C4.3 Optical actuation and control *George K Knopf* 1453
C4.4 Optical to electrical energy conversion: solar cells *Tom Markvart* 1479
C4.5 Medical applications of photonics *Tuan Vo-Dinh* 1501

THE ART OF PRACTICAL OPTOELECTRONICS *Roel Bates*

C5 The art of practical optoelectronic systems *Anthony E Smart* 1519

A1.1
An introduction to optoelectronics

Alan Rogers

A1.1.1 Objective

In this chapter, we shall take a quite general look at the nature of photons and electrons (and of their interactions) in order to gain a familiarity with their overall properties, insofar as they bear upon our subject. Clearly it is useful to acquire this 'feel' in general terms before getting immersed in some of the finer detail which, whilst very necessary, does not allow the inter-relationships between the various aspects to remain sharply visible. The intention is that the familiarity acquired by reading this chapter will facilitate an understanding of the other chapters in the book.

Our privileged vantage point for the modern views of light has resulted from a laborious effort by many scientists over many centuries, and a valuable appreciation of some of the subtleties of the subject can be obtained from a study of that effort. A brief summary of the historical development is our starting point.

A1.1.2 Historical sketch

The ancient Greeks speculated on the nature of light from about 500 BC. The practical interest at that time centred, inevitably, on using the sun's light for military purposes; and the speculations, which were of an abstruse philosophical nature, were too far removed from the practicalities for either to have much effect on the other.

The modern scientific method effectively began with Galileo (1564–1642), who raised experimentation to a properly valued position. Prior to his time experimentation was regarded as a distinctly inferior, rather messy activity, definitely not for true gentlemen. (Some reverberations from this period persist, even today!) Newton was born in the year in which Galileo died, and these two men laid the basis for the scientific method which was to serve us well for the following three centuries.

Newton believed that light was corpuscular in nature. He reasoned that only a stream of projectiles, of some kind, could explain satisfactorily the fact that light appeared to travel in straight lines. However, Newton recognized the difficulties in reconciling some experimental data with this view, and attempted to resolve them by ascribing some rather unlikely properties to his corpuscles; he retained this basic corpuscular tenet, however.

Such was Newton's authority, resting as it did on an impressive range of discoveries in other branches of physics and mathematics, that it was not until his death (in 1727) that the views of other men such as Euler, Young and Fresnel began to gain their due prominence. These men believed that light was a wave motion in a 'luminiferous aether', and between them they developed an impressive theory which well explained all the known phenomena of optical interference and diffraction. The wave theory rapidly gained ground during the late 18th and early 19th centuries.

The final blow in favour of the wave theory is usually considered to have been struck by Foucault (1819–1868) who, in 1850, performed an experiment which proved that light travels more slowly in water than in air. This result agreed with the wave theory and contradicted the corpuscular theory.

For the next 50 years the wave theory held sway until, in 1900, Planck (1858–1947) found it mathematically convenient to invoke the idea that light was emitted from a radiating body in discrete packets, or 'quanta', rather than continuously as a wave. Although Planck was at first of the opinion that this was no more than a mathematical trick to explain the experimental relation between emitted intensity and wavelength, Einstein (1879–1955) immediately grasped the fundamental importance of the discovery and used it to explain the photoelectric effect, in which light acts to emit electrons from matter: the explanation was beautifully simple and convincing. It appeared, then, that light really did have some corpuscular properties.

In parallel with these developments, there were other worrying concerns for the wave theory. From early in the 19th century its protagonists had recognized that 'polarization' phenomena, such as those observed in crystals of Iceland spar, could be explained if the light vibrations were transverse to the direction of propagation. Maxwell (1831–1879) had demonstrated brilliantly (in 1864), by means of his famous field equations, that the oscillating quantities were electric and magnetic fields.

However, there arose persistently the problem of the nature of the 'aether' in which these oscillations occurred and, in particular, how astronomical bodies could move through it, apparently without resistance. A famous experiment in 1887, by Michelson and Morley, attempted to measure the velocity of the earth with respect to this aether, and consistently obtained the result that the velocity was zero. This was very puzzling in view of the earth's known revolution around the sun. It thus appeared that the medium in which light waves propagate did not actually exist!

The null result of the aether experiment was incorporated by Einstein into an entirely new view of space and time, in his two theories of relativity: the special theory (1905) and the general theory (1915). Light, which propagates in space and oscillates in time, plays a crucial role in these theories.

Thus physics arrived (ca. 1920) at the position where light appeared to exhibit both particle (quantum) and wave aspects, depending on the physical situation. To compound this duality, it was found (by Davisson and Germer in 1927, after a suggestion by de Broglie in 1924) that electrons, previously thought quite unambiguously to be particles, sometimes exhibited a wave character, producing interference and diffraction patterns in a wave-like way.

The apparent contradiction between the pervasive wave-particle dualities in nature is now recognized to be the result of trying to picture all physical phenomena as occurring within the context of the human scale of things. Photons and electrons appear to behave either as particles or as waves to us only because of the limitations of our modes of thought. We have been conditioned to think in terms of the behaviour of objects such as sticks, stones and waves on water, the understanding of which has been necessary for us to survive, as a species, at our particular level of things.

In fact, the fundamental atomic processes of nature are not describable in these same terms and it is only when we try to force them into our more familiar framework that apparent contradictions such as the wave–particle duality of electrons and photons arise. Electrons and photons are neither waves nor particles but are entities whose true nature is somewhat beyond our conceptual powers. We are very limited by our preference (necessity, almost) for having a mental picture of what is going on.

Present-day physics with its gauge symmetries and field quantizations rarely draws any pictures at all, but that is another story…

A1.1.3 The wave nature of light

In 1864, Clerk Maxwell was able to express the laws of electromagnetism known at that time in a way which demonstrated the symmetrical interdependence of electric and magnetic fields. In order to complete the symmetry he had to add a new idea: that a changing electric field (even in free space) gives rise to a magnetic field. The fact that a changing magnetic field gives rise to an electric field was already well known, as Faraday's law of induction.

Since each of the fields could now give rise to the other, it was clearly conceptually possible for the two fields mutually to sustain each other, and thus to propagate as a wave. Maxwell's equations formalized these ideas and allowed the derivation of a wave equation.

This wave equation permitted free-space solutions which corresponded to electromagnetic waves with a defined velocity; the velocity depended on the known electric and magnetic properties of free space, and thus could be calculated. The result of the calculation was a value so close to the known velocity of light as to make it clear that light could be identified with these waves, and was thus established as an electromagnetic phenomenon.

All the important features of light's behaviour as a wave motion can be deduced from a detailed study of Maxwell's equations. We shall limit ourselves here to a few of the basic properties.

If we take Cartesian axes Ox, Oy, Oz (figure A1.1.1) we can write a simple sinusoidal solution of the free-space equations in the form:

$$\begin{aligned} E_x &= E_0 \ \exp[\mathrm{i}(\omega t - kz)] \\ H_y &= H_0 \ \exp[\mathrm{i}(\omega t - kz)]. \end{aligned} \tag{A1.1.1}$$

These two equations describe a wave propagating in the Oz direction with electric field (E_x) oscillating sinusoidally (with time t and distance z) in the xz plane and the magnetic field (H_y) oscillating in the yz plane. The two fields are orthogonal in direction and have the same phase, as required by the form of Maxwell's equations: only if these conditions obtain can the two fields mutually sustain each other. Note also that the two fields must oscillate at right angles to the direction of propagation, Oz. Electromagnetic waves are transverse waves.

The frequency of the wave described by equation (A1.1.1) is given by:

$$f = \frac{\omega}{2\pi}$$

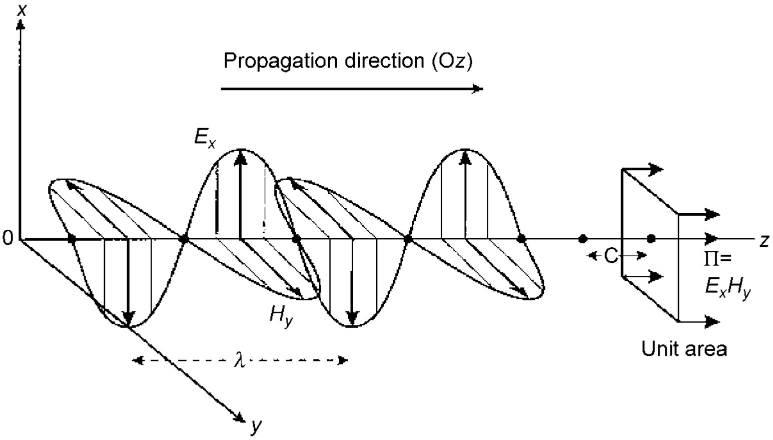


Figure A1.1.1. Sinusoidal electromagnetic wave.

and its wavelength by:

$$\lambda = \frac{2\pi}{k}$$

where ω and k are known as the angular frequency and propagation constant, respectively. Since f intervals of the wave distance λ pass each point on the Oz axis per second, it is clear that the velocity of the wave is given by:

$$c = f\lambda = \frac{\omega}{k}.$$

The free-space wave equation shows that this velocity should be identified as follows:

$$c_0 = \frac{1}{(\varepsilon_0\mu_0)^{1/2}} \tag{A1.1.2}$$

where ε_0 is a parameter known as the electric permittivity, and μ_0 the magnetic permeability, of free space. These two quantities are coupled, independently of equation (A1.1.2), by the fact that both electric and magnetic fields exert mechanical forces, a fact which allows them to be related to a common force parameter, and thus to each other. This 'force-coupling' permits a calculation of the product $\varepsilon_0\mu_0$ which, in turn, provides a value for c_0, using equation (A1.1.2). (Thus Maxwell was able to establish that light in free space consisted of electromagnetic waves.)

We can go further, however. The free-space symmetry of Maxwell's equations is retained for media which are electrically neutral and which do not conduct electric current. These conditions obtain for a general class of materials known as dielectrics; this class contains the vast majority of optical media. In these media the velocity of the waves is given by:

$$c = (\varepsilon\varepsilon_0\mu\mu_0)^{-1/2} \tag{A1.1.3}$$

where ε is known as the relative permittivity (or dielectric constant) and μ the relative permeability of the medium. ε and μ are measures of the enhancement of electric and magnetic effects, respectively, which are generated by the presence of the medium. It is, indeed, convenient to deal with new parameters for the force fields, defined by:

$$\boldsymbol{D} = \varepsilon\varepsilon_0\boldsymbol{E}$$

$$\boldsymbol{B} = \mu\mu_0\boldsymbol{H}$$

where $\boldsymbol{D}$ is known as the electric displacement and $\boldsymbol{B}$ the magnetic induction of the medium. More recently they have come to be called the electric and magnetic flux densities, respectively.

The velocity of light in the medium can (from equation (A1.1.3)) also be written as

$$c = \frac{c_0}{(\varepsilon\mu)^{1/2}} \tag{A1.1.4}$$

where c_0 is the velocity of light in free space, with an experimentally determined value of $2.997925 \times 10^8\,\mathrm{m\,s^{-1}}$. For most optical media of any importance we find that $\mu \approx 1$, $\varepsilon > 1$ (hence the name 'dielectrics'). We have already noted that they are also electrical insulators. For these, then, we may write equation (A1.1.4) in the form:

$$c \approx \frac{c_0}{\varepsilon^{1/2}} \tag{A1.1.5}$$

and note that, with $\varepsilon > 1$, c is smaller than c_0. Now the refractive index, n, of an optical medium is a measure of how much more slowly light travels in the medium compared with free space, and is defined by:

$$n = \frac{c_0}{c}$$

and thus

$$n \approx \varepsilon^{1/2}$$

from equation (A1.1.5).

This is an important relationship because it connects the optical behaviour of the optical medium with its atomic structure. The medium provides an enhancement of the effect of an electric field because that field displaces the atomic electrons from their equilibrium position with respect to the nuclei; this produces an additional field and thus an effective magnification of the original field. The detailed effect on the propagation of the optical wave (which, of course, possesses an electric component) will be considered in chapter A1.2 but we can draw two important conclusions immediately. First, the value of the refractive index possessed by the material is clearly dependent upon the way in which the electromagnetic field of the propagating wave interacts with the atoms and molecules of the medium. Second, since there are known to be resonant frequencies associated with the binding of electrons in atoms, it follows that we expect ε to be frequency dependent. Hence, via equation (A1.1.5), we expect n also to be frequency dependent. The variation of n (and thus of optical wave velocity) with frequency is a phenomenon known as optical dispersion and is very important in optoelectronic systems, not least because all practical optical sources emit a range of different optical frequencies, each with its own value of refractive index.

We turn now to the matters of momentum, energy and power in the light wave. The fact that a light wave carries momentum and energy is evident from a number of its mechanical effects, such as the forced rotation of a conducting vane in a vacuum when one side is exposed to light (figure A1.1.2). A simple wave picture of this effect can be obtained from a consideration of the actions of the electric and magnetic fields of the wave when it strikes a conductor. The electric field will cause a real current to flow

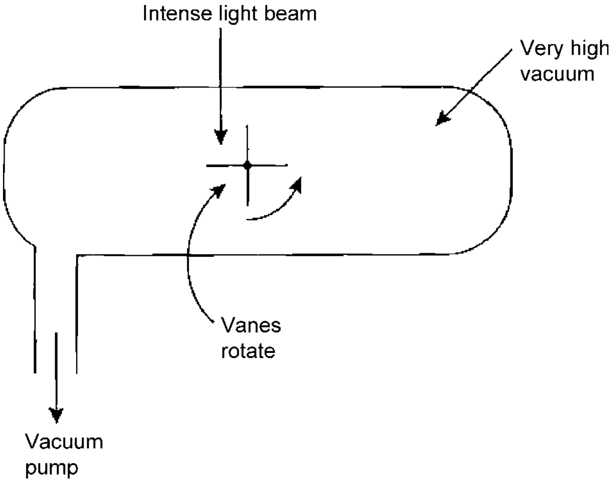


Figure A1.1.2. Force exerted by light falling on a conducting vane.

in the conductor (it acts on the 'free' electric charges in the conductor) in the direction of the field. This current then comes under the influence of the orthogonal magnetic field of the wave. A current-carrying conductor in a magnetic field which lies at right angles to the current flow experiences a force at right angles to both the field and the current (motor principle) in a direction which is given by Fleming's left-hand rule (this direction turns out to be, fortunately, the direction in which the light is travelling!). Hence the effect on the conductor is equivalent to that of energetic particles striking it in the direction of travel of the wave; in other words, it is equivalent to the transport of momentum and energy in that direction.

We can take this description one stage further. The current is proportional to the electric field and the force is proportional to the product of the current and the magnetic field, hence the force is proportional to the product of electric and magnetic field strengths. The flow of energy, that is the rate at which energy is transported across unit area normal to the direction of propagation, is just equal to the vector product of the two quantities;

$$\boldsymbol{\Pi} = \boldsymbol{E} \times \boldsymbol{H}$$

(the vector product of two vectors gives another vector whose amplitude is the product of the amplitudes of the two vectors multiplied by the sine of the angle between their directions (in this case $\sin 90° = 1$) and is in a direction orthogonal to both vectors, and along a line followed by a right-handed screw rotating from the first to the second vector. Vectors often combine in this way so it is convenient to define such a product).

Clearly, if $\boldsymbol{E}$ and $\boldsymbol{H}$ are in phase, as for an electromagnetic wave travelling in free space, then the vector product will always be positive. $\boldsymbol{\Pi}$ is known as the Poynting vector. We also find that, in the case of a propagating wave, $\boldsymbol{E}$ is proportional to $\boldsymbol{H}$, so that the power across unit area normal to the direction of propagation is proportional to the square of the magnitude of either $\boldsymbol{E}$ or $\boldsymbol{H}$. The full quantitative relationships will be developed in later chapters, but we may note here that this means that a measurement of the power across unit area, a quantity known as the intensity of the wave (sometimes the 'irradiance') provides a direct measure of either $\boldsymbol{E}$ or $\boldsymbol{H}$ (figure A1.1.1). This is a valuable inferential exercise since it enables us, via a simple piece of experimentation (i.e. measurement of optical power) to get a handle on the way in which the light will interact with atomic electrons, for example. This is because, within the atom, we are dealing with electric and magnetic fields acting on moving electric charges.

The units of optical intensity, clearly, will be watts metre^{-2}.

A1.1.4 Polarization

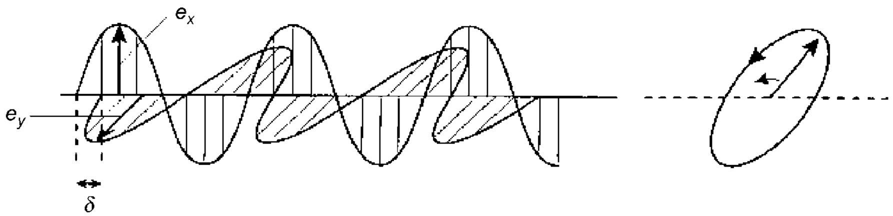
The simple sinusoidal solution of Maxwell's wave equation for $\boldsymbol{E}$ and $\boldsymbol{H}$ given by equation (A1.1.1) is only one of an infinite number of such solutions, with $\boldsymbol{E}$ and $\boldsymbol{H}$ lying in any direction in the xy plane, and with ω taking any value greater than zero.

It is customary to fix attention on the electric field for purposes of general electromagnetic wave behaviour, primarily because the effect of the electric field on the electrical charges within atoms tends to be more direct than that of the magnetic field. But the symmetry which exists between the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields of the electromagnetic wave means that conclusions arrived at for the electric field have close equivalence for the magnetic field. It is simply convenient only to deal with one of them rather than two.

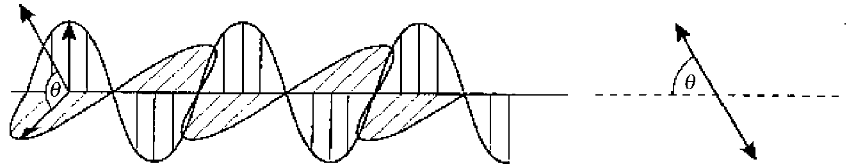
Suppose that we consider two orthogonal electric field components of a propagating wave, with the same frequency but differing phases (figure A1.1.3(*a*)):

$$E_x = e_x \cos(\omega t - kz + \delta_x)$$

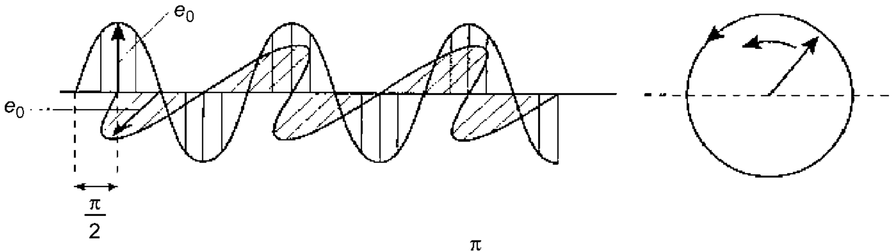
$$E_y = e_y \cos(\omega t - kz + \delta_y).$$



(a) Elliptical polarization: e_x, e_y, $\delta(=\delta_y-\delta_x)$



(b) Linear polarization: e_x, e_y, $\delta=m\pi$



(c) Circular polarization: $e_x=e_y=e_0$, $\delta=(2m+1)\,\frac{\pi}{2}$

Figure A1.1.3. Linear and circular polarization as special cases of elliptical polarization.

From figure A1.1.3 we can see that the resulting electric field will rotate as the wave progresses, with the tip of the resulting vector circumscribing (in general) an ellipse. The same behaviour will be apparent if attention is fixed on one particular value of z and the tip of the vector is now observed as it progresses in time. Such a wave is said to be elliptically polarized. (The word ‘polarized’, being associated, as it is, with the separation of two dissimilar poles, is not especially appropriate. It derives from the attempt to explain crystal-optical effects within the early corpuscular theory by regarding the light corpuscles as rods with dissimilar ends, and it has persisted.) Of notable interest are the special cases where the ellipse degenerates into a straight line or a circle (figure A1.1.3(b) and (c)). These are known as linear and circular polarization states, respectively, and their importance lies not least in the fact that any given elliptical state can be resolved into circular and linear components, which can then be dealt with separately. The light will be linearly polarized, for example, when either e_x or $e_y=0$, or when $\delta_y-\delta_x=m\pi$. It will be circularly polarized only when $e_x=e_y$ and $\delta_y-\delta_x=(2m+1)\pi/2$, where m is a positive or negative integer: circular polarization requires the component waves to have equal amplitude and to be in phase quadrature. A sensible, identifiable polarization state depends crucially on the two components maintaining a constant phase and amplitude relationship. All of these ideas are further developed in chapter A2.1.

The polarization properties of light waves are important for a number of reasons. For example, in crystalline media, which possess directional properties, the propagation of the light will depend upon its polarization state in relation to the crystal axes. This fact can be used either to probe crystal structure or to control the state of the light via the crystal. Furthermore, the polarization state of the light can provide valuable insights into the restrictions imposed on the electrons which gave rise to it.

Wherever there is directionality (i.e. the properties of the medium vary with spatial direction) in the medium in which the light is travelling, the polarization state of the light will interact with it and this is an extremely useful attribute with a number of important applications.

A1.1.5 The electromagnetic spectrum

Hitherto in this chapter we have dealt with optical phenomena in fairly general terms and with symbols rather than numbers. It may help to fix ideas somewhat if some numbers are quoted.

The wave equation allows single-frequency sinusoidal solutions and imposes no limit on the frequency. Furthermore, the equation is still satisfied when many frequency components are present simultaneously. If they are phase-related then the superposition of the many waveforms provides a determinable time function via the well known process of Fourier synthesis. If the relative phases of the components are varying with time, then we have 'incoherent' light; if the spread of frequencies in this latter case exceeds the bandwidth of the optical detector (e.g. the human eye) we sometimes call it 'white' light.

The electromagnetic spectrum is shown in figure A1.1.4. In principle, it ranges from (almost) zero frequency to infinite frequency. In practice, since electro-magnetic wave sources cannot be markedly smaller than the wave-length of the radiation which they emit, the range is from the very low frequency ($\sim 10^3$ Hz) radio waves ($\lambda \sim 300$ km) to the very high frequency ($\sim 10^{20}$ Hz) gamma radiation, where the limit is that of the very high energy needed for their production.

The most energetic processes in the universe are those associated with the collapse of stars and galaxies (supernovae, black holes), and it is these which provide the radiation of the highest observable frequencies.

Visible radiation lies in the range 400–700 nm (1 nm $= 10^{-9}$ m), corresponding to a frequency range of 7.5×10^{14}–4.3×10^{14} Hz. The eye has evolved a sensitivity to this region as a result of the fact that it corresponds to a broad maximum in the spectral intensity distribution of sunlight at the earth's surface: survival of the species is more likely if the sensitivity of the eye lies where there is most light!

The infrared region of the spectrum lies just beyond 700 nm and is usually taken to extend to about 300 000 nm ($\equiv$ 300 μm; we usually switch to micrometres for the infrared wavelengths, in order to keep the number of noughts down).

The ultraviolet region lies below 400 nm and begins at about 3 nm. Clearly, all of these divisions are arbitrary, since the spectrum is continuous.

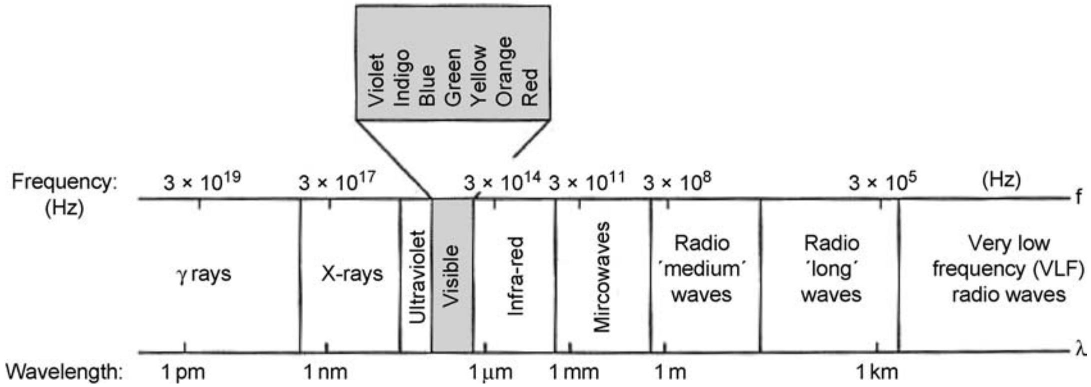


Figure A1.1.4. The electromagnetic spectrum.

It is worth noting that the refractive index of silica (an important optical material) in the visible range is ~ 1.47, so the velocity of light at these wavelengths in this medium is close to $2 \times 10^8\,\mathrm{m\,s^{-1}}$. Correspondingly, at the given optical frequencies, the wavelengths in the medium will be $\sim 30\%$ less than those in air, in accordance with the relation: $\lambda = c/f$. (The frequency will remain constant.)

It is important to be aware of this wavelength change in a material medium, since it has a number of noteworthy consequences which will be explored in chapter A1.2.

A1.1.6 Emission and absorption processes

So far, in our discussions, the wave nature of light has dominated. However, when we come to consider the relationships between light and matter, the corpuscular, or (to use the modern word 'particulate'), nature of light begins to dominate. In classical (i.e. pre-quantum theory) physics, atoms were understood to possess natural resonant frequencies resulting from a conjectured internal elastic structure. These natural resonances were believed to be responsible for the characteristic frequencies emitted by atoms when they were excited to oscillate by external agencies. Conversely, when the atoms were irradiated with electromagnetic waves at these same frequencies, they were able to absorb energy from the waves, as with all naturally resonant systems interacting with sympathetic driving forces. This approach seemed to provide a natural and reasonable explanation of both the emission and absorption spectral characteristics of particular atomic systems.

However, it was soon recognized that there were some difficulties with these ideas. They could not explain why, for example, in a gas discharge, some frequencies were emitted by the gas and yet were not also absorbed by it in its quiescent state; neither could they explain why the energy with which electrons were emitted from a solid by ultraviolet light (in the photoelectric effect) depends not on the quantity of absorbed light energy but only on the light's frequency.

We now understand the reasons for these observations. We know that atoms and molecules can exist only in discrete energy levels. These energy levels can be arranged in order of ascending value: E_0, E_1, $E_2 \cdots E_m$ (where m is an integer) and each such sequence is characteristic of a particular atom or molecule. The highest energy level corresponds to the last below the level at which the atom becomes ionized (i.e. loses an electron).

Fundamental thermodynamics (classical!) requires that under conditions of thermal equilibrium the number, N_i, of atoms having energy E_i is related to the number N_j having energy E_j by the Boltzmann relation:

$$\frac{N_i}{N_j} = \exp\left[-\frac{(E_i - E_j)}{kT}\right]. \tag{A1.1.6}$$

Here k is Boltzmann's constant ($1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$) and T is the absolute temperature.

The known physics now states that light of frequency ν_{ij} can be either emitted or absorbed by the system only if they corresponds to a difference between two of the discrete energy levels, in accordance with the relation

$$h\nu_{ij} = E_i - E_j$$

where h is Planck's quantum constant ($6.626 \times 10^{-34}\,\mathrm{J\,s}$). The more detailed interpretation is that when, for example, an atom falls from an energy state E_j to E_i, a 'particle' of light with energy $h\nu_{ij}$ is emitted. This 'quantum' of light is called the photon; we use the symbol ν to denote frequency rather than f (or $\omega/2\pi$) to emphasize that light is now exhibiting its particulate, rather than its wave, character.

Thus, the relationship between light and matter consists of the interaction between atoms (or molecules) and photons. An atom either absorbs/emits a single photon, or it does not. There is no intermediate state.

The classical difficulties to which reference was made earlier are now resolved. First, some lines are emitted from a gas discharge which are not present in the absorption spectrum of the quiescent gas because the energetic conditions in the discharge are able to excite atoms to high energy states from which they can descend to some lower states; if these states are not populated (to any measurable extent) in the cold gas, however, there is no possibility of a corresponding incoming frequency effecting these same transitions and hence being absorbed. Second, for an incoming stream of photons, each one either interacts or does not interact with a single atom. If the photon energy is higher than the ionization energy of the atom then the electron will be ejected. The energy at which it is ejected will be the difference between the photon energy and the ionization energy. Thus, for a given atom, the ejection energy will depend only on the frequency of the photon.

Clearly, in light/matter interactions, it is convenient to think of light as a stream of photons. If a flux of p photons of frequency ν crosses unit area in unit time then the intensity of the light (defined by the Poynting vector) can be written

$$I = ph\nu. \tag{A1.1.7}$$

It is not difficult to construct any given quantity in the photon approach which corresponds to one within the wave approach. However, there does still remain the more philosophical question of reconciling the two approaches from the point of view of intellectual comfort. The best that can be done at present is to regard the wave as a ‘probability’ function, where the wave intensity determines the probability of ‘finding’ a photon in a given volume of space. This is a rather artificial stratagem which does, however, work very well in practice. It does not really provide the intellectual comfort which we seek, but that, as has been mentioned earlier, is a fault of our intellect, not of the light!

Finally, it may be observed that, since both the characteristic set of energy levels and the return pathways from an excited state are peculiar to a particular atom or molecule, it follows that the emission and/or absorption spectra can be used to identify and quantify the presence of species within samples, even at very small partial concentrations. The pathway probabilities can be calculated from quantum principles, and this whole subject is a sophisticated, powerful and sensitive tool for quantitative materials analysis. It is not, however, within the scope of this chapter.

A1.1.7 Photon statistics

The particulate view of light necessitates the representation of a light flux as a stream of photons ‘guided’ by an electromagnetic wave. This immediately raises the question of the arrival statistics of the stream.

To fix ideas let us consider the rate at which photons are arriving at the sensitive surface of a photodetector.

We begin by noting that the emission processes which gave rise to the light in the first place are governed by probabilities, and thus the photons are emitted, and therefore also arrive, randomly. The light intensity is a measurable, constant (for constant conditions) quantity which, as we have noted, is to be associated with the arrival rate p according to equation (A1.1.7), i.e. $I = ph\nu$. It is clear that p refers to the mean arrival rate averaged for the time over which the measurement of I is made. The random arrival times of the individual particles in the stream imply that there will be statistical deviations from this mean, and we must attempt to quantify these if we are to judge the accuracy with which I may be measured.

To do this we begin with the assumption that atoms in excited states emit photons at random when falling spontaneously to lower states. It is not possible to predict with certainty whether any given excited atom will or will not emit a photon in a given, finite time interval. Added to this there is the knowledge that for light of normal, handleable intensities, only a very small fraction of the atoms in the source material will emit photons in sensible detection times. For example, for a He–Ne laser with an output power of 5 mW, only 0.05% of the atoms will emit photons in 1 s.

Thus we have the situation where an atom may randomly either emit or not emit a photon in a given time, and the probability that it will emit is very small: this is the prescription for Poisson statistics, i.e. the binomial distribution for very small event probability (see, for example, Kaplan, 1981).

Poisson statistics is a well-developed topic and we can use its results to solve our photon arrival problem.

Suppose that we have an assemblage of N atoms and that the probability of any one of them emitting a photon of frequency ν in time τ is q, with $q \ll 1$.

Clearly, the most probable number of photons arriving at the detector in time τ will be Nq and this will thus also be the average (or mean) number detected, the average being taken over various intervals of duration τ. But the actual number detected in any given time τ will vary according to Poisson statistics, which state that the probability of detecting r photons in time τ is given by (Kaplan, 1981):

$$P_r = \frac{(Nq)^r}{r!}\exp(-Nq).$$

Hence the probability of receiving no photons in τ is $\exp(-Nq)$, and of receiving two photons is $[(Nq)^2/2!]\exp(-Nq)$ and so on.

Now the mean optical power received by the detector clearly is given by:

$$P_{\mathrm{m}} = \frac{Nqh\nu}{\tau} \tag{A1.1.8}$$

and P_{m} is the normally-measured quantity. Hence equation (A1.1.8) allows us to relate the mean of the distribution to a measurable quantity, i.e.

$$Nq = \frac{P_{\mathrm{m}}\tau}{h\nu} = \frac{P_{\mathrm{m}}}{h\nu B}$$

where B is the detector bandwidth ($B = 1/\tau$). Now we need to quantify the spread of the distribution in order to measure the deviation from the mean, and this is given by the standard deviation which, for the Poisson distribution, is the square root of the mean. Thus the deviation of the arrival rate is

$$D = (Nq)^{1/2} = \left(\frac{P_{\mathrm{m}}}{h\nu B}\right)^{1/2}.$$

This deviation will comprise a ‘noise’ on the measured power level and will thus give rise to a noise power

$$P_{\mathrm{noise}} = \left(\frac{P_{\mathrm{m}}}{h\nu B}\right)^{1/2}\frac{h\nu}{\tau} = (P_{\mathrm{m}}h\nu B)^{1/2}.$$

Thus the signal-to-noise ratio will be given by

$$\mathrm{SNR} = \frac{P_{\mathrm{m}}}{P_{\mathrm{noise}}} = \left(\frac{P_{\mathrm{m}}}{h\nu B}\right)^{1/2}.$$

This is an important result. It tells us what is the fundamental limit on the accuracy with which a given light power can be measured. We note that the accuracy increases as $(P_\mathrm{m}/h\nu)^{1/2}$, and it is thus going to be poor for low rates of photon arrival. This we would expect intuitively, since the 'granular' nature of the process will inevitably be more noticeable when there are fewer photons arriving in any given time. It will also be poor for large optical frequencies, since this means more energy per photon, and thus fewer photons for a given total light energy. Again the 'granular' nature will be more evident. For good SNR, therefore, we need large powers and low frequencies. Radio wave fluxes from nearby transmitters are easy to measure accurately, gamma rays from a distant galaxy are not.

Finally, it should be remembered that the above conclusions only apply strictly when the probability q is very small. For the very intense emissions from powerful lasers ($\sim 10^6\ \mathrm{W\,m^{-2}}$, say) a substantial proportion of the atoms will emit photons in a typical detection time. Such light is sometimes classed as non-Poissonian (or sub-Poissonian) for reasons which will now be clear.

A1.1.8 The behaviour of electrons

Our subject is optoelectronics, and so far we have been concerned almost exclusively with just one half of it: with optics. The importance of our subject derives from the powerful interaction between optics and electronics, so we should now evidently gain the necessary equivalent familiarity with electronics, to balance our view. We shall, therefore, now look at the general behaviour of electrons.

A free electron is a fundamental particle with negative electrical charge (e) equal to 1.602×10^{-19} C and mass (m) equal to 9.11×10^{-31} kg.

All electrical charges exert forces on all other charges and, for any given charge, q, it is convenient to summarize the effect of all other charges by defining the electric field, $\boldsymbol{E}$, via the value of the force $\boldsymbol{F}_\mathrm{E}$ which the field exerts on q:

$$\boldsymbol{F}_\mathrm{E} = q\boldsymbol{E}.$$

A magnetic field exerts no force on a stationary charge. When the charge moves with velocity v with respect to a magnetic field of induction $\boldsymbol{B}$, however, the force on the charge is given by

$$\boldsymbol{F}_\mathrm{B} = q(v \times \boldsymbol{B})$$

where $v \times \boldsymbol{B}$ denotes the vector product of v and $\boldsymbol{B}$, so that the force is orthogonal to both the vectors v and $\boldsymbol{B}$. Of course, a uniformly moving charge comprises an electrical current, so that $v \times \boldsymbol{B}$ also describes the force exerted by a magnetic field on a current-carrying conductor. The two forces are combined in the Lorentz equation:

$$\boldsymbol{F} = q(\boldsymbol{E} + v \times \boldsymbol{B}) \tag{A1.1.9}$$

which also is a full classical description of the behaviour of the electron in free space, and is adequate for the design of many electron beam devices (such as the cathode-ray tube of television sets) where the electron can be regarded as a particle of point mass subject to known electromagnetic forces.

If an electron (or other electrical charge) is accelerating, then it comprises an electric current which is varying with time. Since a constant current is known to give rise to a constant magnetic field, a varying current will give rise to a varying magnetic field, and this, as we have seen, will give rise in turn to an electric field. Thus an accelerating electron can be expected to radiate electromagnetic waves. For example, in a dipole antenna (figure A1.1.5) the electrons are caused to oscillate sinusoidally along a conducting rod. The sinusoidal oscillation comprises accelerated motion, and the antenna radiates radio waves.

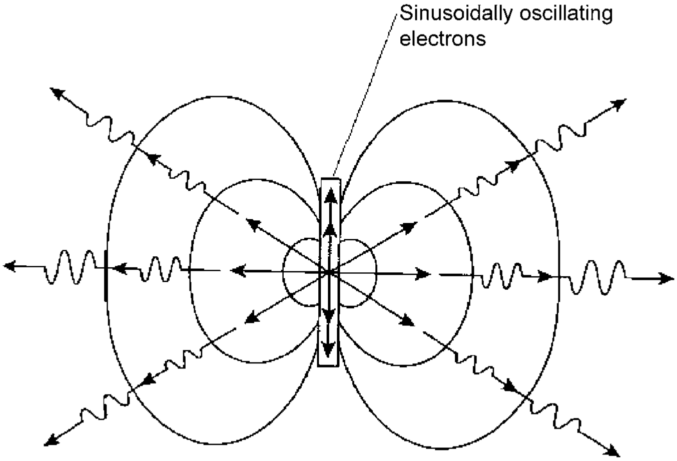


Figure A1.1.5. The radiating dipole.

However, the electron also itself exhibits wave properties. For an electron with momentum p there is an associated wavelength λ given by

$$\lambda = \frac{h}{p}$$

which is known as the de Broglie wavelength, after the Frenchman who, in 1924, suggested that material particles might exhibit wave properties. (The suggestion was confirmed by experiment in 1927.) Here h is, again, the quantum constant.

The significance assigned to the wave associated with the electron is just the same as that associated with the photon: the intensity of the wave (proportional to the square of the amplitude) is a measure of the probability of finding an electron in unit volume of space. The wave is a ‘probability’ wave. The particle/wave duality thus has perfect symmetry for electrons and photons. One of the direct consequences of this duality, for both entities, is the uncertainty principle, which states that it is fundamentally impossible to have exact knowledge of both momentum and position simultaneously, for either the photon or the electron. The uncertainty in knowledge of momentum, Δp, is related to the uncertainty in position, Δx, by the expression:

$$\Delta p \Delta x \approx \frac{h}{2\pi}.$$

There is a corresponding relation between the uncertainty in the energy (ΔE) of a system and the length of time (Δt) over which the energy is measured:

$$\Delta E \Delta t \approx \frac{h}{2\pi}.$$

The interpretation in the wave picture is that the uncertainty in momentum can be related to the uncertainty in wavelength, i.e.

$$p = \frac{h}{\lambda}$$

so that

$$\Delta p = \frac{-h\Delta\lambda}{\lambda^2}$$

and hence

$$\Delta x = \frac{h}{2\pi\Delta p} = \frac{\lambda^2}{2\pi\Delta\lambda}.$$

Hence, the smaller the range of wavelengths associated with a particle, the greater is the uncertainty in its position (Δx). In other words the closer is the particle's associated wave function to a pure sine wave, having constant amplitude and phase over all space, the better is its momentum known: if the momentum is known exactly, the particle might equally well be anywhere in the universe!

The wave properties of the electron have many important consequences in atomic physics. The atomic electrons in their orbits around the nucleus, for example, can only occupy those orbits which allow an exact number of wavelengths to fit into a circumference: again, the escape of electrons from the atomic nucleus in the phenomenon of β-radioactivity is readily explicable in terms of the 'tunnelling' of waves through a potential barrier. But probably the most important consequence of these wave properties, from the point of view of our present discussions, is the effect they have on electron behaviour in solids, for the vast majority of optoelectronics is concerned with the interaction between photons and electrons in solid materials. We shall, therefore, need to look at this a little more closely.

The primary feature which solids possess compared with other states of matter (gas, liquid, plasma) is that the atoms or molecules of which they are composed are sufficiently close together for their electron probability waves to overlap. Indeed, it is just this overlap which provides the interatomic bonding strength necessary to constitute a solid material, with its resistance to deformation.

When two identical atoms, with their characteristic set of energy levels, come close enough for their electronic wave functions (i.e. their waves of probability) to overlap, the result is a new set of energy levels, some lower, some higher than the original values (figure A1.1.6). The reason for this is analogous to what happens in the case of two identical, coupled, mechanical resonant systems, say two identical pendulums, which are allowed to interact by swinging them from a common support rod (figure A1.1.7). If one pendulum is set swinging, it will set the other one in motion, and eventually the second will be swinging with maximum amplitude while the first has become stationary. The process then reverses back to the original condition and this complete cycle recurs with frequency f_B. The system, in fact, possesses

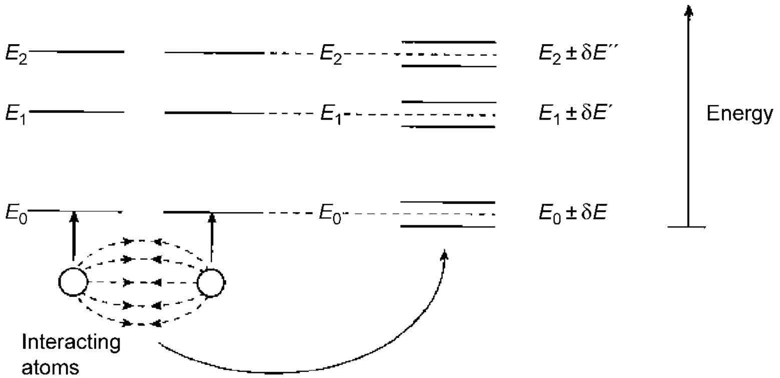


Figure A1.1.6. Splitting of energy levels for two interacting atoms.

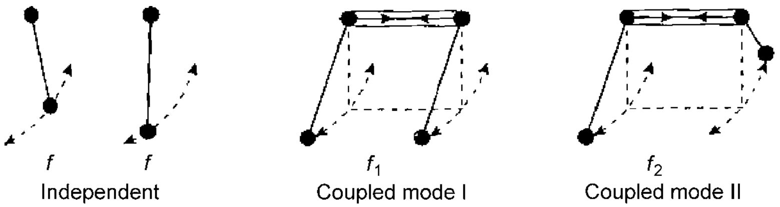


Figure A1.1.7. Interacting pendulums.

two time-independent normal modes: one is where both pendulums are swinging with equal amplitude and are in phase; the other with equal amplitudes in anti-phase. If these two frequencies are f_1 and f_2 we find

$$f_1 - f_2 = f_B$$

and the frequency of each pendulum when independent, f, is related to these by

$$f_1 = f + \frac{1}{2} f_B$$

$$f_2 = f - \frac{1}{2} f_B$$

i.e. the original natural frequency of the system, f, has been replaced under interactive conditions by two frequencies, one higher (f_1) and one lower (f_2) than f.

It is not difficult to extend these ideas to atoms and to understand that when a large number of identical atoms is involved, a particular energy level becomes a band of closely spaced levels. Hence, in a solid, we may expect to find bands separated by energy gaps, rather than discrete levels separated by gaps; and that, indeed, is what is found.

The band structure of solids allows us to understand quite readily the qualitative differences between the different types of solid known as insulators, conductors and semiconductors, and it will be useful to summarize these ideas.

We know from basic atomic physics that electrons in atoms will fill the available energy states in ascending order, since no two electrons may occupy the same state: electrons obey the Pauli exclusion principle. This means that, at the absolute zero of temperature, for N electrons the lowest N energy states will be filled (figure A1.1.8(*a*)). At a temperature above absolute zero the atoms are in thermal motion and some electrons may be excited to higher states, from which they subsequently decay, setting up a dynamic equilibrium in which states above the lowest N have a mean level of electron occupation. The really important point here is that it is only those electrons in the uppermost states which can be excited to higher levels, since it is only for those states that there are empty states within reach (figure A1.1.8(*b*)). This fact has crucial importance in the understanding of solid state behaviour. The electrons are said to have a Fermi–Dirac distribution among the energy levels at any given temperature, rather than the Maxwell–Boltzmann distribution they would have if they were not constrained within the solid, and which is possessed by freely-moving gas molecules, for example.

Consider now the energy band structure shown in figure A1.1.9(*a*). Here the lower band is filled with electrons and there is a large energy gap before the next allowable band, which is empty. The available electrons thus have great difficulty in gaining any energy. If an electric field is applied to this solid it would have very little effect on the electrons, since in order to move in response to the force exerted by the field, they would need to gain energy from it, and this they cannot do, since they cannot

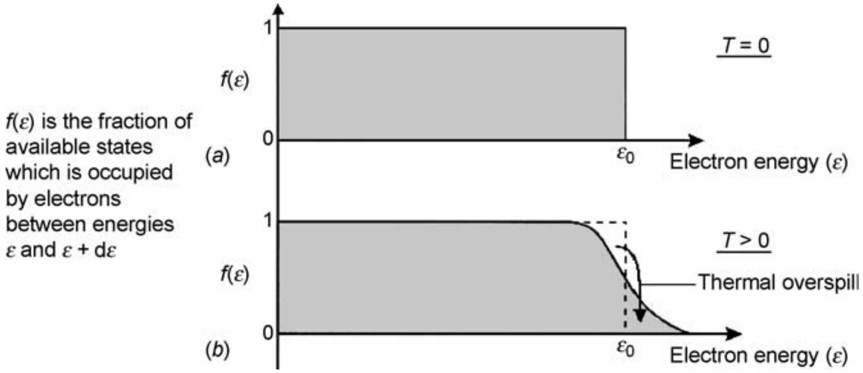


Figure A1.1.8. The Fermi–Dirac distribution for electrons in solids.

jump the gap. Hence the electrons do not move; no current flows in response to an applied voltage; the material is an insulator.

Consider now the situation in figure A1.1.9(*b*). Here the upper band is only half full of electrons. (The electrons in this band will be those in the outer reaches of the atom, and hence will be those responsible for the chemical forces between atoms, i.e. they are valency electrons. Consequently, the highest band to contain any electrons is usually called the valence band.) The situation now is quite different from the previous one. The electrons near the top of the filled levels now have an abundance of unfilled states within easy reach and can readily gain energy from external agencies, such as an applied electric field. Electric currents thus flow easily in response to applied voltages; the material is a metallic conductor.

The third case figure A1.1.9(*c*) looks similar to the first, the only difference being that the gap between the filled valence band and the next higher unoccupied band is now much smaller. As a result, a relatively small number of electrons can be excited into the higher band (known as the conduction band) by thermal collisions and, once there, they can then move freely in response to an applied electric field. Hence there is a low level of conductivity and the material is a semiconductor; more specifically it is an intrinsic semiconductor. It is clear that the conductivity will rise with temperature since more energetic thermal collisions will excite more electrons into the conduction band. This is in contrast to metallic conductors in which the conductivity falls with temperature (owing to greater interference from the more strongly vibrating fixed atoms). There is a further important feature in the behaviour of intrinsic semiconductors. When an electron is excited from the valence band into the conduction band it leaves behind an unfilled state in the valence band. This creates mobility in the valence band, for electrons there

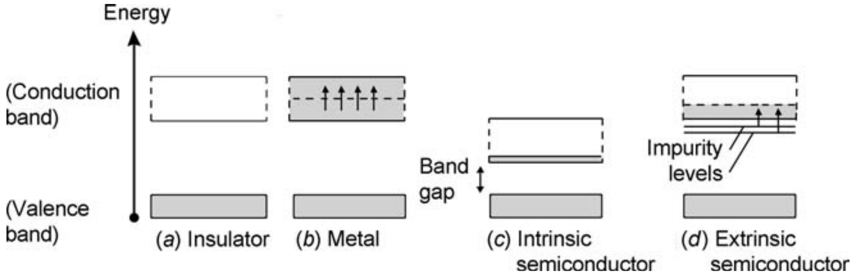


Figure A1.1.9. Energy-level schematic for the three main classes of solid ($T > 0$).

which previously had no chance of gaining energy can now do so by moving into the empty state, or hole, created by the promotion of the first electron. Further, the valence electron which climbs into the hole, itself leaves behind another hole which can be filled in turn. The consequence of all this activity is that the holes appear to drift in the opposite direction to the electrons when an electric field is applied, and thus they are behaving like positive charges. (This is hardly surprising because they are created by the absence of negative charge.) Hence we can view the excitation of the electron to the conduction band as a process whereby an electron/hole pair is created, with each particle contributing to the current which flows in response to an applied voltage.

Finally, we come to another very important kind of semiconductor. It is shown in figure A1.1.9(*d*). Here we note that there are discrete energy levels within the region of energy 'forbidden' to states, the gap between bands. These are due to intruders in the solid, to 'impurities'.

To understand what is going on, consider solid silicon. Silicon atoms are tetravalent (i.e. have a valency of four), and in the solid state they sit comfortably in relation to each other in a symmetrical three-dimensional lattice (figure A1.1.10). Silicon is an intrinsic semiconductor with an energy gap between the filled valence band and the empty (at absolute zero) conduction band of 1.14 eV. (An electron volt is the kinetic energy acquired by an electron in falling through a potential of 1 V, and is equal to 1.6×10^{-9} J.) The Boltzmann factor (equation (A1.1.6)) now allows us to calculate that only about one in 10^{20} electrons can reach the conduction band at room temperature; but since there are of order 10^{24} electrons per cm^3 in the material as a whole, there are enough in the conduction band to allow it to semiconduct.

Suppose now that some phosphorus atoms are injected into the silicon lattice. Phosphorus is a pentavalent (valency of five) atom, so it does not sit comfortably within the tetravalent (valency of four) silicon structure. Indeed, it finds itself with a spare valence electron (it has five as opposed to silicon's four) after having satisfied the lattice requirements. This electron is loosely bound to the phosphorus atom and thus is easily detached from it into one of the conduction band states, requiring little energy for the excitation. Effectively, then, the electron sits in a state close to the conduction band (as shown in figure A1.1.9(*d*)) and, depending on the density of phosphorus atoms (i.e. the 'doping' level), can provide significantly greater conductivity than is the case for pure silicon. Such impurity-doped materials are called extrinsic semiconductors.

As the impurity we chose donated an electron to the conduction band (as a result of having one spare) it is called an n-type semiconductor, since it donates negative charge carriers. Conversely, we could have doped the silicon with a tervalent (valency of three) element, such as boron, in which case it would sit in the lattice in need of an extra electron, since it has only three of its own. The consequence of

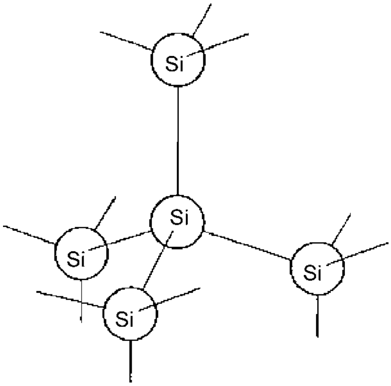


Figure A1.1.10. Structure of silicon lattice.

this will be that a neighbouring silicon valence electron can easily be excited into that vacant state, leaving a positive hole in the valence band as a consequence. This hole now enhances the electrical conductivity, leading to p-type ('positive carrier') semiconductivity. It is now easy to understand why 'pentavalent' elements are said to give rise to 'donor' energy levels and 'tervalent' elements to 'acceptor' levels (in silicon).

There are several reasons why extrinsic semiconductors are so important. The first is that the level of conductivity is under control, via the control of the dopant level. The second is that p-type and n-type materials can be combined with great versatility in a variety of devices having very valuable properties, the most notable of which is the transistor: many thousands of these can now be integrated on to electronic chips.

We are now in a position to understand, in general terms, the ways in which photons can interact with electrons in solids.

Consider again the case of an intrinsic semiconductor, such as silicon, with a band-gap energy E_g. Suppose that a slab of the semiconductor is irradiated with light of frequency ν such that

$$h\nu > E_g.$$

It is clear that the individual photons of the incident light possess sufficient energy to promote electrons from the valence band to the conduction band, leaving behind positive 'holes' in the valence band. If a voltage is now applied to the slab, a current, comprised of moving electrons and holes, will flow in response to the light: we have a *photo*conductor. Moreover, the current will continue to flow for as long as the electron can remain in the conduction band, and that includes each electron which will enter the slab from the cathode whenever one is taken up by the anode. Hence the number of electrons and holes collected by the electrodes per second can far exceed the number of photons entering the slab per second, provided that the lifetime of the carriers is large. In silicon the lifetime is of the order of a few milliseconds (depending on the carrier density) and the electron/photon gain can be as large as 10^4. However, this also means that the response time is poor, and thus photoconductors cannot measure rapid changes in light level (i.e. significant changes in less than a few milliseconds).

Small band-gap materials such as indium antimonide must be used to detect infrared radiation since the corresponding photon energy is relatively small. An obvious difficulty with a narrow band gap is that there will be a greater number of thermally excited carriers, and these will constitute a noise level; hence these infrared detectors usually must be cooled for satisfactory performance, at least down to liquid nitrogen temperatures (i.e. $< 77\,\mathrm{K}$)

In order to increase the speed with which the photoconduction phenomenon can be used to make measurements of light level, we use a device consisting of a combination of n- and p- type semiconductor materials. The two types of material are joined in a 'pn junction' which forms a 'photodiode' (figure A1.1.11). In this case the electron/hole pairs created by the incident photons drift in the electric field across the junction, thus giving rise to a measurable current as before; but each is quickly annihilated at the boundaries of the junction by an abundance of oppositely-charged carriers which combine with them. The reduced recombination time leads to a fast response and, with careful design, responses in times of order tens of picoseconds may be achieved. These pn photodiodes, in addition to being fast, are compact, rugged, cheap and operate at low voltage. They are not generally as sensitive as photoconductive devices, however, since they do not allow 'gain' in the way described for these latter devices (unless used in an 'avalanche' mode, of which more in later chapters).

The pn detection process can also be used in reverse, in which case the device becomes a light emitter. For this action electrons are injected into the pn junction by passing a current through it using, now, 'forward' bias. The electrons combine with holes in the region of transition between p and n materials, and, in doing so, release energy. If conditions are arranged appropriately this energy is in

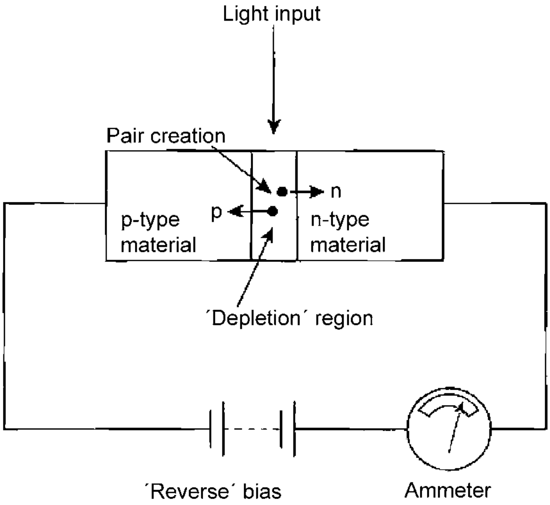


Figure A1.1.11. Schematic view of p–n junction photodiode.

the form of photons and the device becomes an emitter of light—a light-emitting diode (LED). Again this has the advantages of ruggedness, compactness, cheapness and low voltage operation. LEDs already are in widespread use.

A1.1.9 Lasers

Finally, in our general view of optoelectronics, we must have a quick glance at the laser, for that is from where it really all derives.

Our subject began (effectively) with the invention of the laser (in 1960) because laser light is superior in so many ways to nonlaser light. In 'ordinary', so-called 'incoherent', sources of light, each photon is emitted from its atom or molecule largely independently of any other, and thus the parameters which characterize the overall emission suffer large statistical variations and are ill-defined: in the case of the laser, this is not so. The reason is that the individual emission processes are correlated, via a phenomenon known as stimulated emission, where a photon which interacts with an excited atom can cause it to emit another similar photon which then goes on to do the same again, etc. This 'coupling' of the emission processes leads to emitted light which has sharply defined properties such as frequency, phase, polarization state and direction, since these are all correlated by the coupling processes. The sharpness of definition allows us to use the light very much more effectively. We can now control it, impress information upon it and detect it with much greater facility than is the case for its more random counterpart. Add to this facility the intrinsic controllability of electrons via static electric and magnetic fields and we have optoelectronics.

A1.1.10 Summary

Our broad look in this chapter at the subject of optoelectronics has pointed to the most important physical phenomena for a study of the subject and has attempted to indicate the nature of the relationships between them in this context.

Of course, in order to practise the art, we need much more than this. We need to become familiar with quantitative relationships and with a much finer detail of behaviour. These the succeeding chapters will provide.

Acknowledgments

This chapter was first published (by the same author) in *Essentials of Optoelectronics*, Chapman and Hall (1997), and is included here with permission.

Further reading

Bleaney B I and Bleaney B 1985 *Electricity and Magnetism* 3rd edn (Oxford: Oxford University Press) (for a readily digestible account of classical electricity and magnetism, including wave properties).

Cajori F 1989 *A History of Physics* (New York: Macmillan) (for those interested in the historical developments).

Chen C-L 1996 *Elements of Optoelectronics and Fibre Optics* (New York: McGraw-Hill) (a treatment of the subject at a more detailed analytical level).

Ghatak A K and Thyagarajaran K 1989 *Optical Electronics* (Cambridge: Cambridge University Press) (for general optoelectronics at a more advanced level than this chapter).

Goldin E 1982 *Waves and Photons, an introduction to Quantum Theory* (New York: Wiley) (for the basics of photon theory).

Kaplan W 1981 *Advanced Mathematics for Engineers* (Reading, MA: Addison Wesley) p 857 (for a good treatment of Poisson statistics).

Pollock C R 1995 *Fundamentals of Optoelectronics* (New York: McGraw-Hill) (a more mathematical approach to the subject).

Richtmeyer F K, Kennard E H and Lauritsen T 1955 *Introduction to Modern Physics* (New York: McGraw-Hill) (for the physical ideas concerning photons and electrons).

Smith F G and King A 2001 *Optics and Photonics* (New York: Wiley) (a good treatment of the optics/photonics interface).

Solymar L and Walsh D 1993 *Lectures on the Electrical Properties of Materials* 5th edn (Oxford: Oxford University Press) (for a clear treatment of general properties of electrical materials).

A1.2

Optical materials

Neil Ross

A1.2.1 Introduction

Optoelectronics is, in essence, concerned with the interactions between light and the electrons within materials through which the light is propagating. This paper reviews the basic solid state physics that is necessary to understand the behaviour of many optoelectronic devices. The emphasis is on the physical models that are used to understand and predict the behaviour of materials. It is assumed that the reader has some knowledge of the basic principles of quantum mechanics, but no attempt will be made to formulate the models in a rigorous mathematical form. Only inorganic materials will be considered in this paper, as polymers and organic materials are considered elsewhere. As the interaction between light and a material is primarily through the electrons, it will be necessary to review the behaviour of electrons in a solid material in some detail, with particular emphasis on semiconductors because of their technological importance.

A1.2.2 Optical properties of some common materials

Before considering the underlying physics of materials for use, it is appropriate to consider the optical properties of some commonly used materials. The two most fundamental of these optical properties are the *transmission window*, i.e. the range of wavelengths over which the material is able to transmit light, and the refractive index.

Refractive index is defined as the ratio of the speed of light in vacuum to its speed in the material. Strictly this is the ratio of the *phase velocities*, rather than the *group velocities*, of the electromagnetic wave but the difference is rarely significant. For most optical materials, which are transparent in the visible region of the spectrum, this ratio has a value that is within the range of about 1.3–1.8. The refractive index of some different types of optical glass for use in the UV, visible and near infrared regions of the spectrum is shown in figure A1.2.1. Much higher values are often found for materials transmitting in the infrared. For example, zinc selenide has a refractive index of about 2.4, silicon about 3.4 and germanium 4.0. Refractive index is important because it determines the reflection and refraction at the boundaries between materials. This makes it possible to produce familiar components, e.g. lenses, and also guided wave devices, e.g. optical fibres.

For some materials (usually crystalline), the refractive index depends on the polarization of the light and on the direction of propagation. Such materials are said to show *birefringence*. Examples of birefringent materials are quartz (SiO_2) and calcite ($CaCO_3$). The maximum difference in refractive index between the two orthogonal linear polarizations is quite small in quartz (0.009), but quite large in calcite (0.17). Isotropic materials, such as most glasses and some crystals, do not normally show

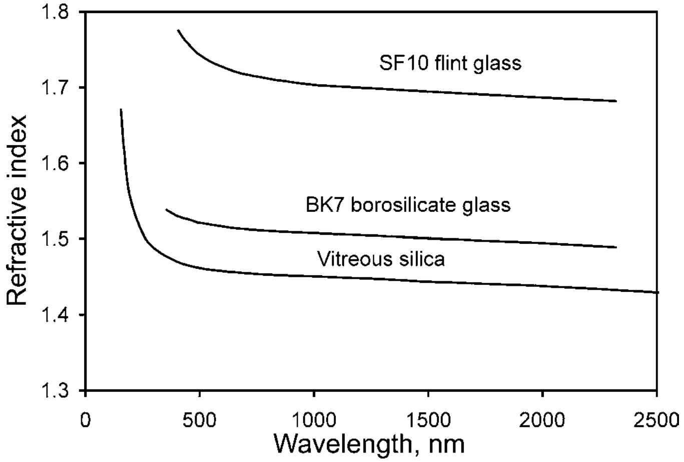


Figure A1.2.1. The refractive index of some optical glasses.

birefringence. However, when an isotropic material is subjected to mechanical strain or to an electric field, which introduces some anisotropy, birefringence may be induced.

The attenuation of light as it travels through a medium is due to *scattering* and *absorption*. Scattering arises from inhomogeneities, possibly on an atomic or molecular scale, and microscopic voids or inclusions in the material. Light is scattered in all directions, rather like the beam of a vehicle headlight in fog. This process leads to an attenuation of the light beam, as the light diffuses out in all directions. Scattering is not usually a limiting factor for many applications of optical materials, because the materials have been chosen for their clarity and optical homogeneity. The limiting factor on transmission is usually absorption. However, for optical fibres, where very high purity, low absorption glasses are used, it is the residual scattering from the microscopic fluctuations in the refractive index that ultimately limits the transmission. These fluctuations in refractive index arise from the thermal density fluctuations that occur in the molten glass. As the glass solidifies, these fluctuations become frozen in.

The other loss mechanism is that the light may be absorbed by the material. Generally, the region of good optical transparency is bounded at the short wavelength end by strong electronic absorption as electrons within the material are excited to higher energy states. At the long wavelength end, it is generally the excitation of molecular vibrations or *phonons* (vibrational or elastic waves) that provides the limit to the region of optical transparency.

Figure A1.2.2 shows the transmission of various optical materials in the visible and infrared parts of the spectrum. The transmission curves show values of the external transmission. That is, the measured transmission includes not only the losses within the material, but also the reflection losses at the two faces of the sample. Often, some of the absorption may be due to impurities in the material. This is the case for the absorption spectrum of UV-grade vitreous silica shown in figure A1.2.2. The strong absorption lines at around 1400, 2300 and 2800 nm are due to hydroxyl (OH) radicals (water). This is a consequence of the method of fabrication of the glass and alternative methods of production lead to much lower absorption in this region, at the cost of increased absorption in the ultra violet. The OH radicals conveniently act to reduce the absorption at short wavelengths by interacting with absorbing *colour centres*, which can otherwise occur in the silica glass. The colour centres are sites where there are nonbridging (i.e. missing) Si–O bonds and they introduce extra electronic states which lead to additional absorption. Hydroxyl (OH) groups may become attached at these sites, reducing the UV absorption.

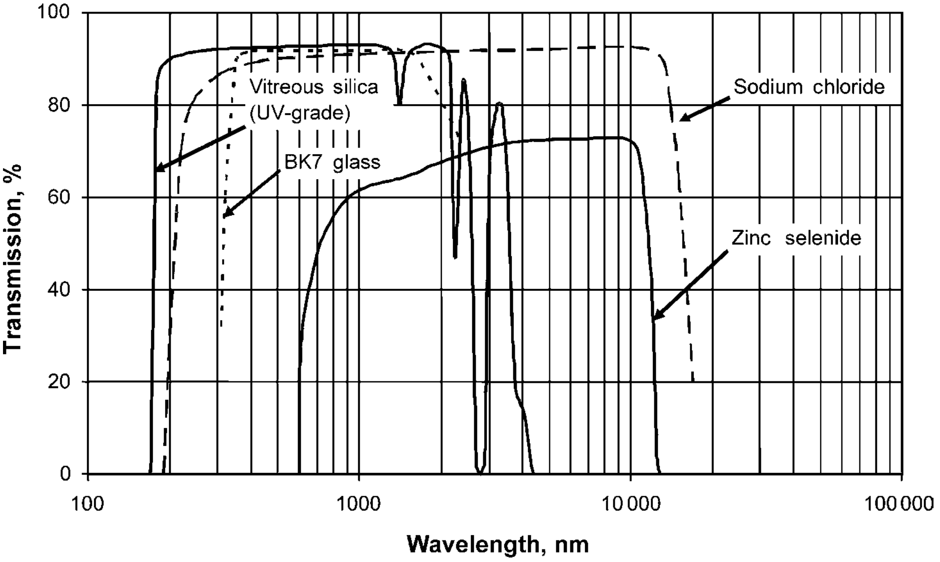


Figure A1.2.2. The external transmission of various optical materials for a 10 mm path length.

A1.2.3 Crystalline and amorphous materials

While glasses are widely used for passive optical components (e.g. lenses, windows and optical fibres), crystalline materials are generally used for many active electro-optic components (light sources, detectors, modulators, etc). Crystalline materials are also the simplest and the best-understood materials. The essential feature of a crystal is that the atoms and molecules are arranged in a periodic pattern. An abstract concept of a periodic structure, the *lattice*, is used to describe the location of the molecules within the crystal [1, chapter 1]. At each so-called *lattice point*, the atoms surrounding the point form the same pattern with the same orientation. Associated with each lattice point is a fixed volume, the *primitive cell*. There are many possible choices for the shape of the primitive cell, but in each case the volume is the same. The group of atoms associated with each lattice point is the *basis*. Together the lattice and the basis define the structure of the crystal. Of particular significance is that they determine its symmetry properties with respect to refraction and reflection of light. These properties are of interest when considering the polarization and nonlinear properties of optical materials.

To illustrate the concept of the lattice and the basis, consider the diamond structure, which is a structure characteristic not only of diamond but also of the elemental semiconductors silicon and germanium. This structure has a face centred cubic structure with lattice points lying at the corners of a cube and also at the centres of the six faces (figure A1.2.3(*a*)). The basis is an atom at a lattice point (0, 0, 0) and a second atom at a point (1/4, 1/4, 1/4), where the units are that of the lattice spacing (figure A1.2.3(*b*)). These combine to give the diamond structure (figure A1.2.3(*c*)), where each atom is bonded to four nearest neighbours. The solid lines indicate the bonds.

The lattice of an ideal single crystal extends throughout the crystal. In practice, defects, which disrupt the regular pattern, are often found in the structure. These defects will, in general, change the physical properties of the material. For electronic or electro-optic devices, it is usually necessary to minimize or eliminate such defects. While some materials, e.g. common salt, are easily identified as crystals, because of the regular shape in which they form, some, e.g. most metals, are not. This is generally because, in the latter, the bulk material is polycrystalline and made up of many small

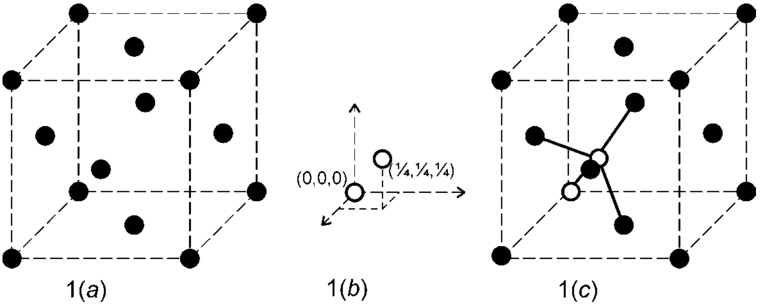


Figure A1.2.3. Diamond crystalline structure: (*a*) face centred cubic lattice; (*b*) basis; (*c*) full structure.

crystallites with varying orientations. These crystallites are not readily visible but can be seen by suitable treatment of the surface, e.g. chemical etching.

Amorphous materials differ from crystals, in that there is no long-range order [1, chapter 17]. Around each atom there may be some semblance of order, in that its nearest neighbours are approximately in the same pattern for all atoms, but this order rapidly decreases as the distance increases. Examples of amorphous materials would be soot, or of more relevance to electro-optics, amorphous silicon. Some of these are soft materials with little mechanical strength. Glasses are also amorphous materials, which may be considered to be composed of one large macromolecule, with strong bonding between atoms, but no long or medium range order or periodicity. They are of considerable importance in optics. Glasses are essentially super-cooled liquids. They have no well-defined melting point, but they progressively soften and become less viscous over a range of temperature. Although there is no long-range order, there is some degree of structure at short distances.

A1.2.4 Atomic bonding

In a crystalline material, strong forces bond the atoms together. This force is primarily due to the interaction of the outer electrons of the atoms (valence electrons). The number of valence electrons determines the chemical properties and the position in the periodic table. Several types of bond are commonly identified.

A1.2.4.1 Ionic bonds

In ionic crystals, two dissimilar atoms are bonded by charge transferred from one atom to the other, leaving a positive ion and a negative ion, with a consequent electrical force bonding the atoms together. A typical example would be common salt, NaCl.

A1.2.4.2 Covalent bonds

In this case, the bonding occurs with electrons being shared among adjacent atoms. A typical example would be silicon. Such bonds are strong, and generally have a well-defined direction, which will determine the structure of the unit cell.

A1.2.4.3 Metallic bonds

In the case of metals, electrons are also shared among atoms, but unlike covalent bonds, the electrons involved are free to move through the whole crystal, giving high electrical and thermal conductivity and,

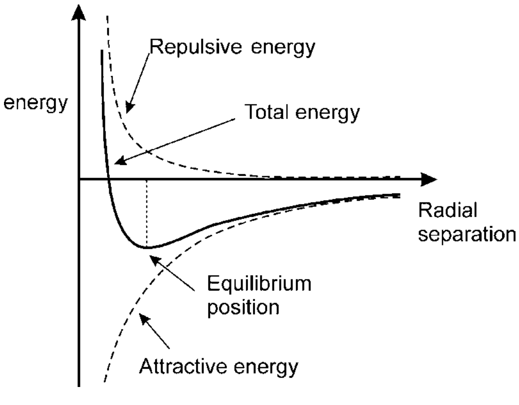


Figure A1.2.4. The binding energy between two atoms.

generally, very low optical transmission. Because of the high conductivity, the reflectivity for light and longer wavelength electro-magnetic waves is usually high.

A1.2.4.4 Van der Waals bonds

These are much weaker bonds arising from a dipole interaction among the atoms. This type of bonding may be important in organic materials.

The attractive forces pull the atoms together, but as they approach more closely, a repulsive force arises due to the interaction of the inner core electrons. Thus, as two isolated atoms are brought together, the potential energy will fall to a minimum and then rise again (figure A1.2.4). The equilibrium separation will be at the location of the minimum. For small displacements from the minimum, the potential energy curve will be approximately quadratic with displacement. Thus, if the atoms are displaced, they will oscillate with simple harmonic motion. This vibrational mode of oscillation generally occurs at frequencies corresponding to the infrared region of the spectrum and may determine the infrared properties of a material.

A1.2.5 The free electron model for metals

A good place to start when looking at the electrons in solid state materials is with metals. While metals are not generally thought of as optical materials, starting the discussion here enables some of the key theoretical concepts to be introduced in a relatively simple way.

In a metal, the valence electrons are free to move through the body of the material. This permits the use of a simple model, which is to treat the electrons as a ‘gas’ of free particles. This provides a good starting point for considering the electronic properties of solid materials, and introduces some of the essential concepts.

The properties of this electron gas are analysed by assuming that the electrons are completely free (no potential energy) up to the boundaries of the material, when the potential becomes infinite [1–5]. Using this ‘electron in a box’ model, the energy states are found by solving the Schrödinger wave equation. Of course, there is no net electric charge. It is assumed that the positive charge of the ionized atoms is uniformly distributed through the whole ‘box’. The model also ignores the interaction among the electrons, and the consequent collective effects. This is a rather odd assumption given the long-range Coulomb forces. Hook and Hall [3, chapter 13], discuss the reasons why this simplified model works.

Inside the box, the electrons can be analysed as simple plane waves, with the walls of the box imposing boundary conditions that require the wave function to be zero outside the box. This requires standing wave solutions. The permitted solutions are characterized by their *wavevector*, $\boldsymbol{k}$, which has a magnitude, k, of $2\pi/\lambda$ where λ is the *de Broglie wavelength* of the electron. Usually k is called the *wavenumber*, although this term is also frequently used for the inverse of the wavelength. The direction of $\boldsymbol{k}$ is the direction in which the electron is moving. Details of the analysis can be found in any good text on solid state physics [1–4].

From this model, it is possible to calculate the *density of states*, $N(E)$:

$$N(E) = \frac{V}{\pi^2 \hbar^3}(2m^3 E)^{1/2} \tag{A1.2.1}$$

where $N(E)\mathrm{d}E$ is the number of states with energy between E and $E + \mathrm{d}E$ for a volume V of the metal.

For a metal at very low temperature, the states will fill up from the bottom. The Pauli exclusion principle requires that each state can be occupied by only one electron. The maximum energy is therefore found by integrating $N(E)$ from zero to E_F and equating this number of states to the number of electrons. The *Fermi energy*, E_F, is given by

$$E_\mathrm{F} = \frac{\hbar^2}{2m}(3\pi^2 n_\mathrm{e})^{2/3} \tag{A1.2.2}$$

where n_e is the number of electrons per unit volume. The value of k corresponding to the Fermi energy forms a spherical surface in three-dimensional k-space. This is the *Fermi surface.*

To give some idea of the magnitude of the Fermi energy, consider potassium, a simple metal for which the free electron theory works well. A mole of potassium has a volume of $4.54 \times 10^{-5}\,\mathrm{m}^2$ and contains 6.0×10^{23} atoms, each atom contributes one electron, so $n_\mathrm{e} = 1.32 \times 10^{28}$ electrons m^{-3}. Substituting this into equation (A1.2.2) gives the Fermi energy as 2.04 eV.

At finite temperatures, electrons will be excited to energies somewhat higher than the Fermi energy and unfilled states will be left at lower energies. The Fermi–Dirac distribution function, $f(E)$, governs the probability that any particular state is occupied by an electron:

$$f(E) = \frac{1}{\mathrm{e}^{(E-\mu)/k_\mathrm{B}T} + 1} \tag{A1.2.3}$$

where k_B is Boltzmann's constant, T the absolute temperature and μ the *chemical potential*. At absolute zero temperature, μ is equal to E_F. At finite temperatures, the value will vary, but only slowly. Provided $k_\mathrm{B}T$ is much less than E_F, μ and E_F may be assumed equal. In figure A1.2.5, $f(E)$ is plotted for a chemical potential of 2 eV, at temperatures of 0, 300 and 600 K. In practice, the transition from 1 to 0 is usually quite sharp since the thermal energy kT is usually much less than μ. The chemical potential is also commonly referred to as the *Fermi level*.

The number of electrons with energies lying between E and $E + \mathrm{d}E$ is given by

$$n(E) = N(E)f(E)\,\mathrm{d}E. \tag{A1.2.4}$$

The energy of the electrons may also be related to their momentum, or wavenumber. Classically, the energy of an electron is given by

$$E = \frac{p^2}{2m}$$

where E is the electron energy, p the electron momentum and m the electron mass. The quantum mechanical equivalent is obtained by replacing p with $\hbar k$. This gives the relationship between energy, E,

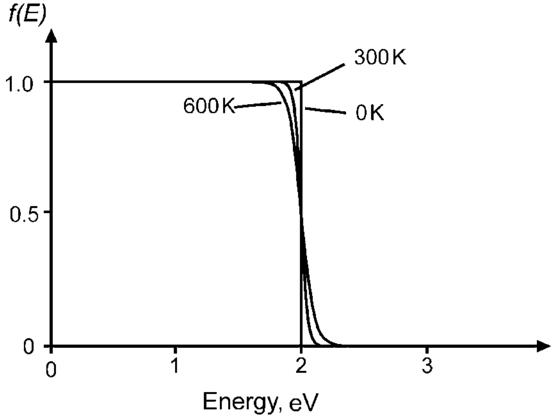


Figure A1.2.5. The Fermi distribution function at three temperatures.

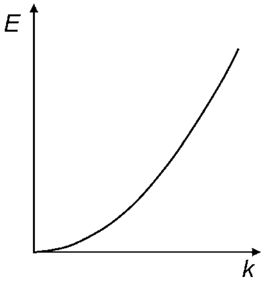


Figure A1.2.6. The parabolic variation of electron energy with k for the free electron model.

and the electron wavenumber, k, for free electrons:

$$E = \frac{\hbar^2 k^2}{2m}. \tag{A1.2.5}$$

Diagrams plotting E against k are useful for understanding band structure, particularly in semiconductors. The E–k diagram for free electrons (figure A1.2.6) is the simplest example.

At zero temperature, with no electric field, the electrons will lie within the Fermi surface in k-space. At finite temperatures, the edge of this sphere will not be distinct but the distribution will still be spherically symmetric. When an electric field is applied to the metal, the electrons will acquire a drift velocity in a direction determined by the field. The net effect is that the whole distribution of electrons in k-space will be shifted slightly, while still maintaining its spherical symmetry. This shift corresponds to a small increase in momentum in the direction of motion and a small increase in total energy.

A1.2.6 Electrons in a periodic lattice

The free electron model has neglected the presence of the atoms associated with the crystal lattice. These atoms will provide a potential that will attract electrons at a large distance from the core of the atom and repel electrons that move so close as to interact with the tightly bound electrons of the atom's core. A one-dimensional representation of the potential along a line through a series of atoms is illustrated in

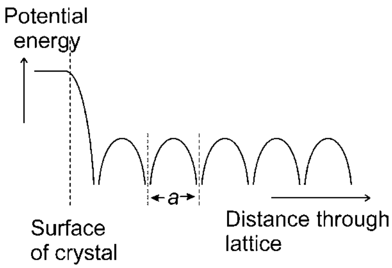


Figure A1.2.7. A representation of the periodic potential experienced by an electron due to the attraction of atomic nuclei.

figure A1.2.7. The most significant feature of the potential is the periodic attractive potential due to the charged atomic cores, separated by a distance a. If the electrons have a de Broglie wavelength equal to the distance between atoms, then strong reflections may be expected. The solution of the Schrödinger wave equation is clearly more complex with this modulated potential.

The *nearly free electron model* provides insight into the behaviour of electrons in a periodic potential [1–5]. Assuming that the interaction of the electrons with the periodic potential is weak, the electrons behave essentially as free electrons, unless their wavelength is close to the separation of the atoms, when travelling electron waves will interact coherently with the periodic potential and will be reflected. For a one-dimensional model of a crystal, with a line of atoms separated by a distance a, this reflection will occur when the electron wavenumber is given by

$$k = n\pi/a \tag{A1.2.6}$$

where n is an integer. The variation of energy with k would be expected to follow that for free electrons (equation (A1.2.5)), except in the region of k given by equation (A1.2.6). This simple model is able to predict that the electrons may only have energy within certain bands, separated by energy gaps.

In order to investigate the band structure, it is necessary to postulate a form for the potential energy and to solve the Schrödinger equation. One such model is the *Kronig–Penney model* [5]. This assumes a simple rectangular model for the potential along the one-dimensional line of atoms (figure A1.2.8). The Schrödinger equation is solved, assuming that the rectangular potential reduces to a series of delta functions (b tends to zero, keeping bV_0 constant). Solutions of the wave equation are not possible for all values of the energy E. Discontinuities in the E–k diagram occur at $k = n\pi/a$, where n is an integer. Away from these values of k, the solution is approximately parabolic, as for the free electron model. This is illustrated in figure A1.2.9(a). As k approaches $k = n\pi/a$, the gradient of the E–k curve approaches zero.

Usually the E–k diagram is modified by mapping all the bands to lie within the range of k between $-\pi/a$ to $+\pi/a$ (referred to as the *first Brillouin zone*), and usually only the positive values of k are

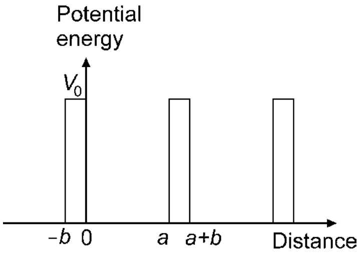


Figure A1.2.8. The periodic potential used in the Kronig–Penney model.

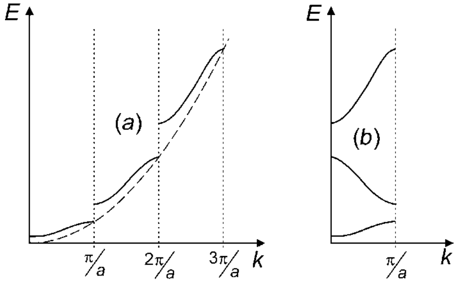


Figure A1.2.9. (*a*) Sketch of the *E–k* diagram obtained using the Kronig–Penney model and (*b*) the same data mapped using the reduced zone scheme.

included, since the negative values give a mirror image. This mapping is referred to as the *reduced zone scheme* [1] and is illustrated in figure A1.2.9(*b*). The mapping is mathematically valid because of the periodicity in k-space of the electron wave functions.

The one-dimensional, nearly free electron model demonstrates that the periodicity of a crystal results in electrons being confined to bands separated by energy gaps. It also enables the dynamics of the electron to be analysed. The free electron equations may be used, provided that the electron mass is replaced by an effective mass. This effective mass takes account of the interaction between the electrons and the lattice (see, for example [3, chapter 4]. The effective mass of an electron is given by

$$m^* = \hbar^2 \left(\frac{\mathrm{d}^2 E}{\mathrm{d}k^2}\right)^{-1}. \tag{A1.2.7}$$

Thus, the effective mass of an electron depends on the curvature of the band and hence will take positive values, in the lower part of a band, negative values in the upper part of the band and will become infinite at some intermediate point.

Another important result from the one-dimensional model is the total number of states in a band. It may be shown [4] that, for a line of atoms of density N_a atoms per unit length and length L, the total number of states is $2N_aL$. Hence, at low temperature and filling up the bands from the lowest energy, all the bands will be full if the atoms have an even number of electrons, and if the number of electrons is odd, the final band will be half full.

Extending the model to three dimensions clearly increases its complexity, but the same general features exist. Obviously, the periodicity will vary with direction through the crystal. The wavenumber must be replaced by the wavevector and the value of $\boldsymbol{k}$ at which the discontinuities occur will now be surfaces in three dimensions. The E–k diagram is still useful, but of course it must now be a plot of E against the wavevector in a particular direction. Frequently, E will be plotted against two different wavevector directions on one diagram, one direction for the positive axis and the other for the negative axis. For real materials, the band structure is much more complex than for the simple one-dimensional model. One particular feature is that, unlike the bands predicted by the simple model, the bands do not necessarily align. The maximum of one band does not necessarily occur at the same k value as the minimum of the next. The significance of this will become apparent shortly.

A1.2.7 Metals, insulators and semiconductors

Conduction of electricity is only possible if there are electrons in a band, having vacant states. A simple way to see this is to recall that, for free electrons, the effect of applying an electric field was to shift

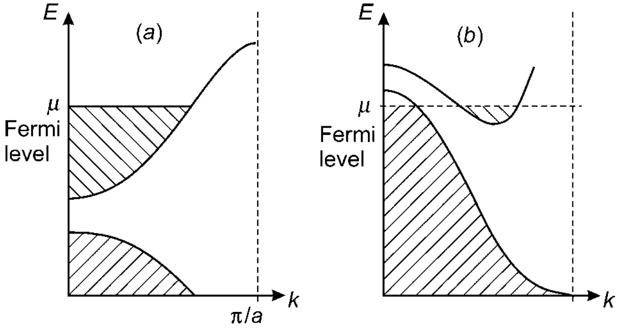


Figure A1.2.10. Schematic showing partially filled bands in a metal: (*a*) a simple metal with partially filled conduction band; (*b*) metal with overlapping bands.

the whole electron distribution slightly to reflect the change in electron momentum when an electric current flows. For a full band, this process is not possible, the distribution of electrons within the band is fixed. A full band cannot contribute to conduction.

A metal must have free electrons; it must therefore have an unfilled band (or bands). For the one-dimensional model discussed earlier, this implies that the number of electrons must be odd, in which case the highest energy band will be only half full. Such a metal will be well described by the free electron model. This simple model works well for the alkali metals, sodium and potassium, which have odd number of electrons. Many metals, however, e.g. magnesium or lead, have an even number of electrons and the simple one-dimensional theory does not work. This is because there is overlap between bands, hence filling up the available energy states from the lowest available states gives two partially filled bands. These two scenarios are illustrated schematically in figures A1.2.10(*a*) and (*b*).

There is a significant difference between the two cases. For the simple metal, the electrons in the conduction band behave very much as predicted by the free electron model, however, in the case of overlapping bands, one band is nearly full and the other nearly empty. Near the Fermi surface, the curvature of the E–k curve for the almost full band will be negative. Hence, the effective mass given by equation (A1.2.7) will also be negative. Rather than considering the effective mass to be negative, it is conventional to introduce the concept of *holes*. In essence, in an almost full band, an electron with an effective negative mass behaves like a positively charged particle with positive effective mass [1, chapter 8, or, 3, chapter 5]. A more physical way to view a hole is that it is the absence of an electron from an atom, leaving a positively charged core, hence the name. The hole is able to move through the lattice as electrons move in the opposite direction, conserving charge. The hole carries positive charge and behaves like a positively charged particle. While these two models for a hole are very different, they are equivalent. In materials where there are two bands contributing to conduction, one almost full and one with only a few of its available states occupied, conduction will effectively be bipolar, with both electrons and holes contributing to the conduction.

If, at low temperature, the electrons exactly fill a number of bands, and there is an energy gap before the next band, then conduction will not be possible and the material is an insulator. At nonzero temperature, the Fermi distribution function (equation (A1.2.3)) will still apply and the probability of electrons in the next band will be nonzero. If the energy gap is large, the number in this higher band will be very small and the material is an insulator. If, however, the energy gap is not so large (~1 eV), then there will be significant excitation at room temperature, and the material will be a semiconductor. The distinction between an insulator and a semiconductor is therefore not clearly defined in this model. However, diamond with a band gap of 5.4 eV would usually be considered an insulator, while silicon with a band gap of 1.17 eV would be a semiconductor. The full band(s) (at low temperature) below the energy gap are referred to as the *valence* band(s), while the band above the energy gap is the *conduction* band.

A1.2.8 Carriers, conduction and doping in semiconductors

In a semiconductor, the Fermi level, or chemical potential, lies within the energy gap. For a pure (intrinsic) semiconductor, the location of the Fermi level is determined by the need to balance the population of electrons in the conduction band with the holes in the valence band. Taking the zero of energy as the top of the valence band, μ is given by [3, chapter 5]

$$\mu = \frac{1}{2}E_G + \frac{3}{4}k_BT\ln\left(\frac{m_h}{m_e}\right) \tag{A1.2.8}$$

where E_G is the energy gap and m_e and m_h the effective electron and hole masses at the bottom of the conduction band and the top of the valence band, respectively. This will generally be close to the centre of the gap, as the second term is small. The density of electrons in the conduction band, n, and holes in the valence band, p, are related by the equation

$$np = n_i^2 = 4\left(\frac{k_BT}{2\pi\hbar^2}\right)^3 (m_h m_e)^{3/2} \exp\left(\frac{E_G}{k_BT}\right). \tag{A1.2.9}$$

In an intrinsic semiconductor $n = p = n_i$.

When an electric field is applied, both the electrons and holes will drift under the influence of the applied field. The drift velocity depends on the strength of the applied field, the carrier density and the rate at which the carriers lose energy to the atoms in the lattice. Generally, it is assumed that the current density $\boldsymbol{j}$ is proportional to the carrier density and to the applied field $\boldsymbol{E}$ (Ohm's law is obeyed). Then the total current density is given by

$$\boldsymbol{j} = (ne\mu_e + pe\mu_h)\boldsymbol{E} \tag{A1.2.10}$$

where μ_e and μ_h are the electron and hole mobilities, respectively, and e the electronic charge. Generally, the electron mobility will be greater than the hole mobility. For silicon, at room temperature, the electron and hole mobilities are about $\mu_e = 1500\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ and $\mu_h = 450\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$. These values may be significantly reduced by impurities in the silicon.

In semiconductors, as in some metals, conduction is due to both the motion of electrons and holes. However, in semiconductors, adding low concentrations of certain impurity atoms (doping) may be used to control both the conductivity and the dominant carriers. In an elemental semiconductor with valency 4, such as silicon or germanium, adding pentavalent impurities such as phosphorous or arsenic will greatly increase the concentration of free electrons giving ‘n type’ material. The impurity atoms can fit into the lattice reasonably well, but the extra electron is not required for the covalent bonding and is only weakly bonded to the *donor* atom. At room temperature, many of these donor atoms are thermally ionized, releasing their electrons into the conduction band where they are free to move. In a similar way, adding trivalent atoms such as boron or aluminium will provide *acceptor* sites, where an electron may be removed from the valence band and trapped. The hole created in the valence band is free to move and will contribute to electrical conduction. Material with this type of doping is ‘p type’. If the electrons or holes from the ionized impurities dominate conduction, the semiconductor is said to be *extrinsic*. If the dominant mechanism by which the electrons and holes are produced is direct thermal excitation of

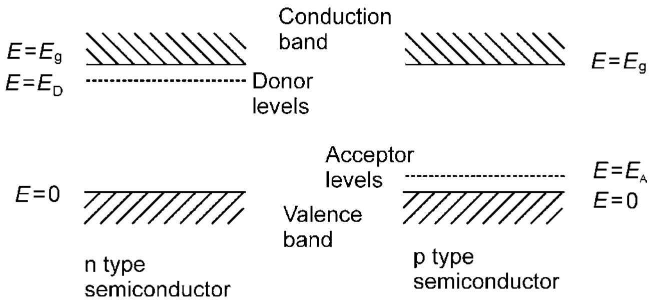


Figure A1.2.11. The location of the donor and acceptor levels within the band gap of a semiconductor.

carriers from the valence to conduction band, then the semiconductor is said to be *intrinsic*. The latter usually exhibits a low conductivity for most semiconductors at room temperature. However, the conductivity of an intrinsic semiconductor increases rapidly with rising temperature.

From the point of view of the energy, the donors or acceptors introduce extra energy levels within the energy gap (figure A1.2.11). The donor or acceptor levels at E_D or E_A lie close to the conduction or valence bands, respectively, and, when ionized, will either donate an electron to the conduction band or accept an electron from the valence band (leaving a hole). Taking the energy of the top of the valence band as zero, and the energy gap as E_g, then the energies E_A and $E_g - E_D$ are typically of order 0.05 eV for the dopants used in silicon.

Equation (A1.2.9) will still hold for an extrinsic semiconductor. Thus, if we consider an n type material, increasing the concentration of donors will increase the number of majority carriers (electrons) and decrease the concentration of minority carriers (holes). For p type materials, the holes will be the majority carriers.

As the temperature is increased from absolute zero, a number of changes occur in a doped semiconductor. At very low temperature, the impurity atoms are not ionized, the carrier concentrations are very small and the Fermi level lies mid-way between E_D and E_G for an n doped material. For a p doped material, the Fermi level will lie at an energy level of $E_A/2$. At very low temperature, the material behaves as an insulator. As the temperature rises, the donors or acceptors are ionized and the majority carrier density increases until all the impurities are ionized (extrinsic region). At the same time, the Fermi level decreases for n doped material, and increases for p doped material. Increasing the temperature still further, direct thermal ionization across the energy gap becomes significant and the semiconductor moves into intrinsic conduction, with the Fermi level moving towards its intrinsic value (equation A1.2.8). These changes are illustrated schematically in figure A1.2.12 for an n doped material.

A1.2.9 The interaction between light and materials

A1.2.9.1 Refraction

Classical electromagnetic theory gives the interaction between a light wave and a medium in terms of the dielectric constant, ε_r, and the conductivity, σ. The solution of Maxwell's equations, for a plane wave of angular frequency ω travelling in the $+z$ direction is of the form [6, chapter 8]

$$E_x = E_{x0} \exp[\mathrm{j}\omega(t - nz/c)] \tag{A1.2.11}$$

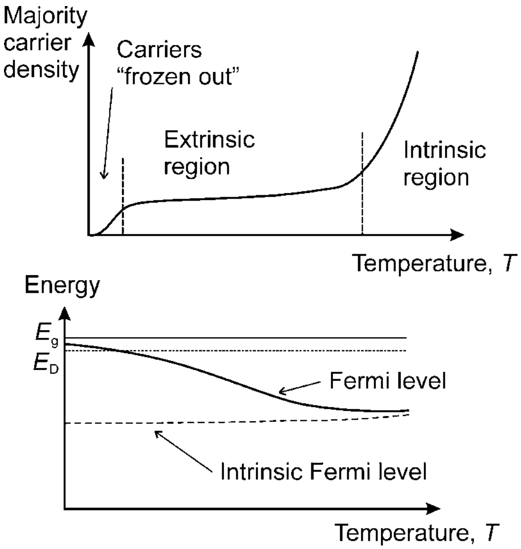

Majority carrier density
Carriers "frozen out"
Extrinsic region
Intrinsic region
Temperature, T
Energy
E_g
E_D
Fermi level
Intrinsic Fermi level
Temperature, T

Figure A1.2.12. A schematic representation of the variation of majority carrier density and Fermi Level in an n type semiconductor.

where E_x is the electric field, n is the (complex) refractive index that is given by:

$$n = n' - jn'' = \sqrt{\mu_r\varepsilon_r - j\left(\frac{\sigma\mu_r}{\omega\varepsilon_0}\right)}. \tag{A1.2.12}$$

In equation (A1.2.12), μ_r is the relative permeability (generally very close to 1 for nonferromagnetic materials), ε_r the dielectric constant, σ the conductivity and ε_0 the permittivity of free space.

Substituting a complex index of refraction into equation (A1.2.11) gives:

$$E_x(\omega) = E_{x0}\exp(-\omega n'' z/c)\exp(j\omega(t - n'z/c))$$

from which it can be seen that the absorption coefficient, α is given by:

$$\alpha = \omega n''/c. \tag{A1.2.13}$$

In a highly conducting material such as a metal, where there are plenty of free electrons, the second term in equation (A1.2.12) will be dominant and n will be of the form

$$n = n' - jn'' = \sqrt{\left(\frac{\sigma\mu_r}{2\omega\varepsilon_0}\right)}(1 - j). \tag{A1.2.14}$$

The absorption coefficient is therefore given by

$$\alpha = \frac{\omega n''}{c} = \left(\frac{\sigma\mu_r\mu_0\omega}{2}\right)^{1/2} = \delta^{-1} \tag{A1.2.15}$$

where c has been replaced by $(\varepsilon_0\mu_0)^{-1/2}$ and δ is the *skin depth*. The skin depth is a measure of the penetration of the field into a conductor. For aluminium, with a resistivity of $6.65 \times 10^{-8}\ \Omega\,\mathrm{m}^{-1}$, the skin depth at a wavelength of 500 nm (green light) is about 3.3 nm, corresponding to an attenuation coefficient $3 \times 10^8\ \mathrm{m}^{-1}$.

For a metal, the conductivity term generally dominates the absorption and the term in ε_r can be neglected. For an insulating material, the conductivity term may be neglected, but not ε_r. The dielectric constant, ε_r, is a measure of the induced polarization of the material by the applied field. The polarization may be due to either the physical alignment of polar molecules, which generally only occurs at frequencies much below optical frequencies, or it may be due to dipole moments in the atoms or molecules induced by the applied field. In dense materials, the local field at a particular molecule or atom is distorted by the adjacent, induced dipoles. Taking account of this distortion (the *Lorentz field*), the Clausius–Mossotti equation may be deduced:

$$\frac{\varepsilon_\mathrm{r} - 1}{\varepsilon_\mathrm{r} + 2} = \frac{1}{3\varepsilon_0} \sum \alpha_i \tag{A1.2.16}$$

where α_i is the polarizability of the i^{th} atom or molecule [3, chapter 9]. The α_i here must not be confused with the absorption coefficient. This relation applies to isotropic materials and needs to be treated with caution. The formula works well for gases but less well for solid state materials where the interaction between the molecular dipoles is stronger.

The simplest model for the polarizability of the atoms is to treat them as classical harmonic oscillators, comprising an electron bound to a fixed atomic core. Using this model α_i will in general be complex. Close to resonance, the imaginary part of α_i, and hence ε_r, will be significant, corresponding to stronger absorption. Well away from the resonance, the polarization will be real and given by

$$\alpha_i = \frac{e^2}{m(\omega_0^2 - \omega^2)}. \tag{A1.2.17}$$

The dc polarizability of each atom is given by $e^2/(m\omega_0^2)$. Combining equations (A1.2.16) and (A1.2.17), and replacing ε_r by the square of the refractive index, the refractive index may be calculated as a function of normalized angular frequency ω/ω_0. The results are plotted in figure A1.2.13, where it has been assumed that the low frequency refractive index is 1.5

The simple model using classical harmonic oscillators clearly has weaknesses. A damping or loss term should be included to avoid the infinity at $\omega = \omega_0$, that can never occur in real materials. Treating the atoms as quantum mechanical oscillators modifies equation (A1.2.17), by introducing a constant, the oscillator strength, f_i. Also, it is not appropriate to treat all the oscillators as having a single resonant frequency. In solid materials, there are usually at least two regions of the frequency spectrum in which such resonances occur. In the infrared, vibrational modes of the lattice (phonons) lead to absorption and dispersion, while at shorter wavelengths, electronic or inter band transitions lead to a further region of strong absorption with associated dispersion.

For visually transparent optical materials, the strong electronic absorption lines occur in the ultraviolet part of the spectrum and, using the simple model, a slowly rising refractive index would be expected over the range of frequencies at which the material is transparent. The refractive index of vitreous silica is shown in figure A1.2.14, as a function of optical frequency. The data shown covers a range of frequencies that correspond to a wavelength range from 2325 to 213.9 nm. As expected from the basic theory, the refractive index rises as the frequency approaches the range in which the silica starts to absorb strongly by electronic transitions. The more rapid fall at the low frequency end of the spectrum occurs as the frequency approaches the near infrared absorption arising from the excitation of vibrational states. In this case, approaching the absorption from the high frequency side, a dip in refractive index would be expected from figure A1.2.13, as is observed in practice.

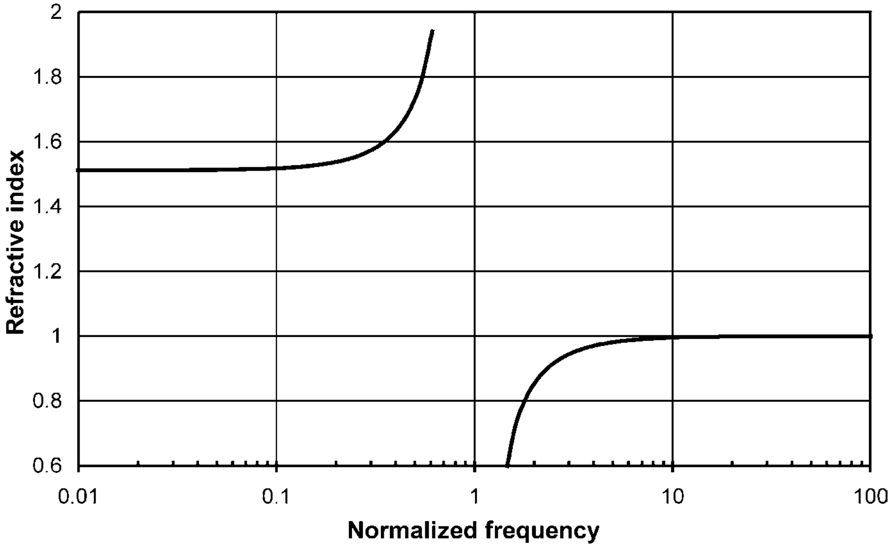


Figure A1.2.13. The refractive index close to resonance as predicted by equations (A1.2.16) and (A1.2.17), with the dc refractive index adjusted to give a refractive index of 1.5 at low frequency.

A1.2.9.2 Absorption and emission

For isolated atoms, light is absorbed when an electron in an atom is excited from one energy state to a higher state (figure A1.2.15(*a*)). Not all transitions are possible as there are selection rules, arising from the need to conserve angular momentum and spin. The interactions only occur if the photon energy matches the difference between the two well-defined states, and the line width is generally narrow. When atoms bond to form larger molecules many additional states are introduced, due to both vibration of the

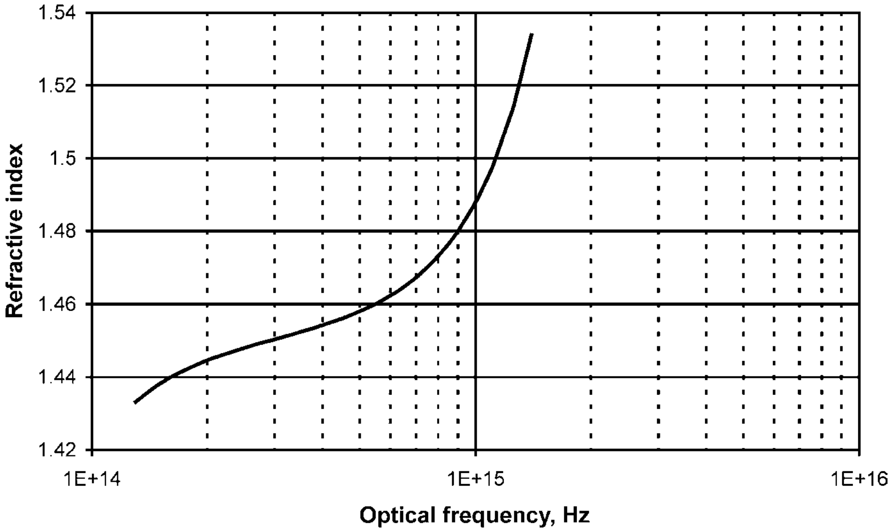


Figure A1.2.14. The refractive index of vitreous silica as a function of optical frequency (data from [7]).

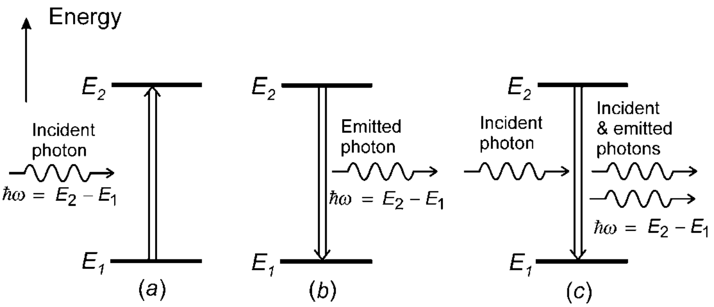

Energy
E_2
E_1
Incident photon
$\hbar\omega = E_2 - E_1$
(a)
E_2
E_1
Emitted photon
$\hbar\omega = E_2 - E_1$
(b)
E_2
E_1
Incident photon
Incident & emitted photons
$\hbar\omega = E_2 - E_1$
(c)

Figure A1.2.15. Energy diagrams illustrating (*a*) absorption, (*b*) spontaneous emission and (*c*) stimulated emission.

bonds between atoms, and to rotation of the molecules. In the condensed state (liquid or solid state), the atoms interact strongly with both near and more distant neighbours and the electronic energy levels of the atoms are spread out into bands and hence the absorption spectrum of liquid and solid state materials generally consists of broad bands, although they may have well-defined edges as discussed later. Absorption also occurs when the optical field excites quantized vibrational waves (phonons), which leads to absorption in the infrared part of the spectrum.

An atom in an excited state may emit a photon as it relaxes to a lower energy state. This emission may occur spontaneously (figure A1.2.15(*b*)) or may be stimulated by another photon of the same energy (figure A1.2.15(*c*)). This process of stimulated emission creates a photon that has the same direction and effective phase as the wave function of the incident photon. Stimulated emission and absorption oppose each other, and the net result is that light may be absorbed or amplified, depending on the relative populations of the atoms in the upper and lower states. For a material close to thermodynamic equilibrium, the population of the lower state will exceed that of the upper state and the material will absorb light. In an optical amplifier, or in a laser that includes an optically amplifying region, the upper state is selectively excited and a *population inversion* occurs, with the population of the upper state exceeding that of a lower state. This provides optical gain.

A1.2.9.3 Fluorescence

Fluorescence is the process by which radiation is absorbed at one wavelength and then re-emitted at another, longer wavelength. Frequently, the excited state generated by the absorption relaxes rapidly by nonradiative processes to a somewhat lower energy state with a longer lifetime, before radiating and relaxing to a lower energy state (possibly the ground state). If the lifetime of the excited state of the radiative transition is long, then the fluorescence may continue for a significant period (microseconds to milliseconds) after the exciting radiation has been cut off. Many materials show fluorescence, some with high efficiency.

A1.2.9.4 Scattering

The final interaction to be considered here is scattering. In this case, the material does not absorb the light, but the incident radiation is scattered into all directions. The simplest form of scattering does not involve any interchange of energy and hence, is called *elastic scattering*. Elastic scattering will occur from any small variations in refractive index of the medium. For example, the scattering of light from a clear transparent glass is usually from tiny regions of the glass, where the refractive index differs from its mean value. These refractive index fluctuations arise from statistical density fluctuations, which occur in the liquid phase and then are frozen in as the glass solidifies. These regions of higher or lower refractive index are much smaller than the wavelength of light. Scattering from particles or regions of varying optical density that have dimensions much less than the wavelength of light ($< \lambda/10$) is called *Rayleigh* scattering. Rayleigh scattering occurs in all materials and in all phases, gas, liquid or solid. The strength of the scattering depends inversely on the fourth power of the wavelength of the light (λ^{-4}). Thus, visible or UV light is scattered much more strongly than infrared.

Raman scattering is, by contrast, an inelastic process in which energy is exchanged with the scattering material. This energy is in the form of molecular vibrations, or optical phonons in the quantum explanation of the effect. The wavelength shift depends on the vibrational energy of the molecule, or the phonon energy. The photon may lose energy (creating a phonon) as it is scattered, decreasing its frequency and increasing its wavelength (figure A1.2.16(*a*)). The decrease in the frequency (usually expressed in wavenumbers) is called the Stokes shift. Alternatively, a photon may gain energy by absorbing energy from a phonon or vibrating molecule (figure A1.2.16(*b*)), giving a shift to a higher frequency (anti-Stokes shift). Raman scattering is normally a weak process, generally several orders of magnitude weaker than Rayleigh scattering, but it is useful as a chemical diagnostic technique. The effect is used in some opto-electronic devices, but in most cases here it is stimulated Raman scattering that is used. Stimulated Raman scattering occurs at high optical intensities, such as may occur in optical fibres, even at moderate power levels. The spontaneous Raman effect may be enhanced if there is an excited state close to resonance with the exciting photon.

The third scattering mechanism is *Brillouin* scattering. This is an inelastic process similar to Raman scattering, except the energy states with which the exchange occurs are lower energy acoustic-mode phonons (sound waves). In a simplified classical description, it can be thought of as light scattered from moving regions of acoustic-wave induced compression and rarefaction, which behave like a diffraction grating. The movement of this ‘grating’ at the acoustic velocity induces a Doppler shift in the scattered light. The frequency shift is much smaller than for Raman scattering and again the effect is generally weak. It should, however, be noted that, as with Raman scattering, in a high optical field, with a long interaction length, the mechanism may lead to a nonlinear effect that results in strong stimulated scattering. This can occur at quite modest power levels (a few milliwatts) if a high coherence laser is launched into a long length of low-loss optical fibre.

The typical spectrum of scattered light from a glass, at moderate frequency resolution, is as shown in figure A1.2.17. The central peak at the excitation wavelength is strong and primarily due to the Rayleigh scattering, but also includes unresolved Brillouin scattering. The frequency shift for the Brillouin scattering is only about 10 GHz, and hence can only be resolved by high-resolution spectroscopic techniques. The relative magnitude of the Stokes and anti-Stokes bands depends on the number of thermally excited phonons, and hence the temperature.

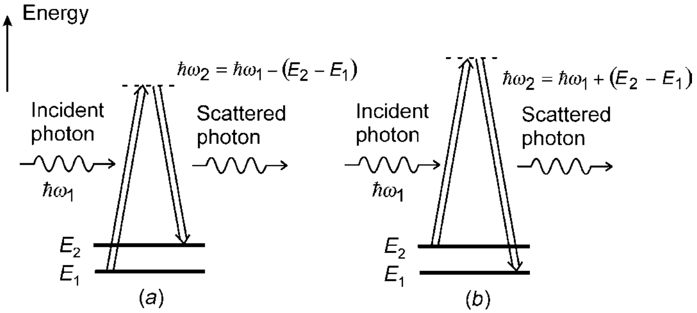


Figure A1.2.16. Energy diagrams illustrating Raman scattering: (*a*) Stokes scattering; (*b*) anti-Stokes scattering.

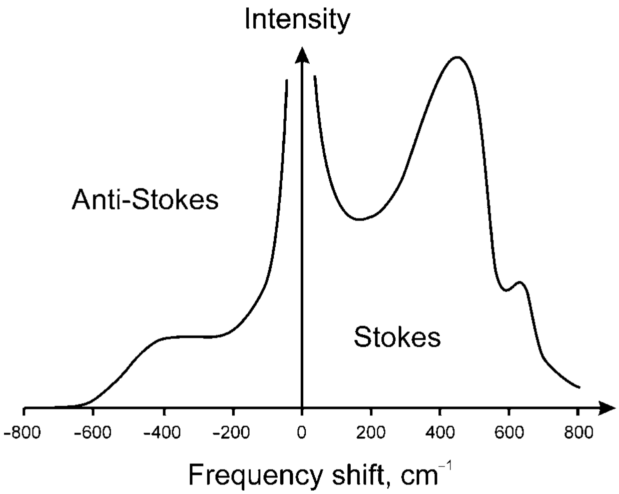


Figure A1.2.17. Raman spectrum at room temperature for a typical silica-based glass.

A1.2.10 The absorption and emission of light by semiconductors

A1.2.10.1 Absorption and photo-conductivity

In the previous discussion on semiconductors, only thermal excitation of electrons between the valence and conduction bands has been considered. Electrons may also be excited by the absorption of electromagnetic radiation. For materials with an energy gap, photons may only excite electrons across the gap if the photon energy exceeds the energy band gap.

$$\hbar\omega \geqslant E_g$$

where ω is the optical angular frequency. There is a very small possibility of multi-photon absorption, but this will be neglected.

When a photon is absorbed, the usual conservation laws (energy, momentum and spin) must all be satisfied. Thus, for a simple transition in which a photon is absorbed and an electron excited, the change in the energy of the electron will be $\Delta E = \hbar\omega$ and the change in the wavevector (momentum) of the electron will be equal to the wavevector of the photon. The wavevector of the photon will, in general, be very small compared with the wavevectors of the electrons. Hence, a simple absorption results in a very small change in $\boldsymbol{k}$. To illustrate this, consider typical values. The lattice spacing for a typical crystal, a, is of order 10^{-10} m, so the wavevector at the limit of the first Brillouin zone has a magnitude (π/a) of 3.1×10^{10}. The wavevector for a photon of wavelength 1 μm (i.e. an energy close to the band gap of silicon) has a magnitude of only 6.3×10^6. It follows that the absorption or emission of a photon will transfer an insignificant amount of momentum to, or from the electrons. To preserve momentum, therefore, the transition is therefore essentially vertical on the E–k diagram.

For a direct gap semiconductor (such as GaAs), the maximum of the valence band lies directly below the minimum of the conduction band (figure A1.2.18(*a*)). In this case, once the photon energy exceeds the band gap, the absorption in the semiconductor will rise rapidly. For an indirect gap semiconductor (such as Si or Ge), there is a significant change of k between the maximum of the valence band and the minimum of the conduction band (figure A1.2.18(*b*)). Transitions between these two points can only occur if a third 'particle' is available to enable momentum (k) to be conserved. This particle is a phonon, a quantized, vibrational excitation of the crystal lattice [1 chapter 8, 4, chapter 8]. The phonon

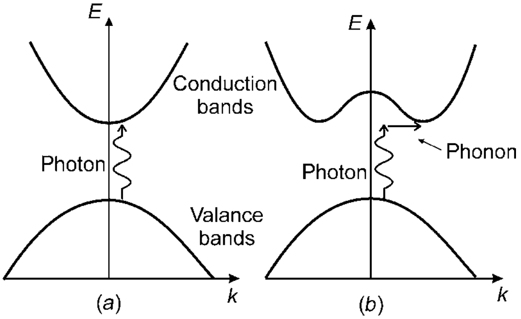


Figure A1.2.18. Diagram illustrating absorption in (*a*) a direct band gap semiconductor and (*b*) an indirect band gap semiconductor.

may be created in the interaction and carry off the momentum and some of the energy or the interaction may be with an existing photon. However, the energy associated with the phonon will be much less than the energy of the photon for the same k value, so most of the energy is transferred to the electron. This *three-body* interaction is much less probable than the simple interaction, where there is no change of electron momentum. Direct band gap semiconductors have an absorption that rises very rapidly near the band edge (photon energy corresponding to the band gap). By contrast, indirect gap semiconductors show a slow rise in absorption until the energy reaches a value that permits direct transitions (figure A1.2.19).

The balance of the excitation rate and the decay rate will determine the steady-state carrier density produced by the radiation and may be written as

$$n_{\mathrm{e}} = n_{\mathrm{h}} = R\tau$$

where τ is the carrier lifetime and R the excitation rate, which will depend on the intensity of the light and the absorption coefficient of the material. The induced carriers will, in turn, lead to increased

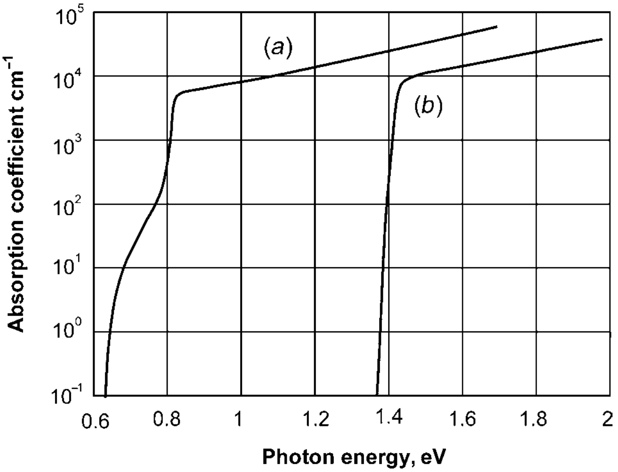


Figure A1.2.19. The absorption of (*a*) germanium and (*b*) gallium arsenide close to the band edge (data taken from [8, 9]).

conductivity (photo-conductivity), which will build up or decay in a time determined by the carrier lifetime.

A1.2.10.2 Emission

If a semiconductor is able to absorb light, exciting electrons from valence band to conduction band, then it will also be possible for light to be emitted by the radiative decay from the conduction band to the valence band. Quantum mechanics shows that the two processes are closely linked. The excited electrons will generally relax very quickly to the bottom of the conduction band, and in the same way, the available holes will float to the top of the valence band. As with absorption, emission of radiation requires that k is conserved, and a high probability of radiative decay requires a simple two-body interaction, hence the electron cannot change its k value. The electrons rapidly decay to the bottom of the conduction band by nonradiative processes, and similarly, the holes rise to the top of the valence band. Hence, in an indirect band gap semiconductor light emission will require the interaction with a phonon of suitable k value. This three-body interaction is, again, of low probability, and recombination is more likely to occur by nonradiative processes. This implies that only direct gap semiconductors can be efficient emitters of light. Obviously, electrons must be excited to the conduction band in order that the semiconductor may emit light. This excitation may be optical, in which case the emission is fluorescence, or it may be due to the injection of minority carriers across a p–n junction.

A1.2.11 Polycrystalline and amorphous semiconductors

In the earlier discussion, it has been assumed that the material is in the form of a perfect crystal. However, as mentioned earlier, crystals usually contain *point defects* and *dislocations*. These are disruptions to the regular lattice and may take a variety of forms [1–3]. Point defects occur when an atom is missing or displaced from its position within the lattice, they may also arise from impurity atoms (as in the deliberate doping of semiconductors). Dislocations are due to imperfections in the lattice which are not localized at a point but extend through the crystal. A simple example is the edge dislocation, when part of a row or plane of atoms is missing, creating stress around the dislocation and leaving *dangling bonds* (figure A1.2.20(*a*)). Dislocations and defects may diffuse through the crystal, especially at elevated temperature. They will modify the band structure introducing extra energy states,

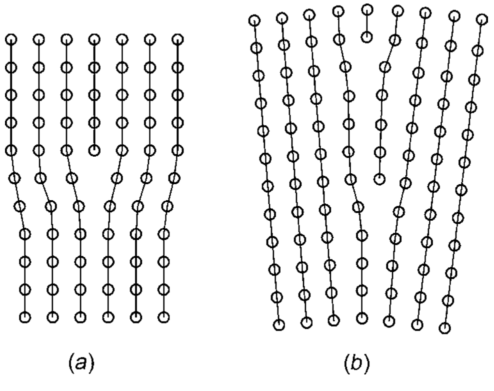


Figure A1.2.20. (*a*) Schematic diagram illustrating an edge dislocation (in cross-section) and (*b*) a small angle grain boundary.

which may trap carriers. Semiconductor devices are usually made from materials, which are as free as possible from dislocations and defects (other than deliberate doping).

In a polycrystalline material, the lattice does not extend unbroken throughout the whole sample, but there are many crystallites or crystal grains with their lattices orientated at different angles. Bonds exist across the grain boundaries, but obviously there is a complex loss of order. For small angles between the lattices at a grain boundary, the boundary may be made up of a row or plane of edge dislocations [1, chapter 20]. This is illustrated schematically for two dimensions in figure A1.2.20(*b*). Grain boundaries will introduce many surface states, which will generally seriously impair the performance of electronic devices. There are, however, cases where the impaired performance may be balanced by a greater need for reduced cost. An example of this is the polycrystalline solar cell, where the lower cost of producing a large area polycrystalline device, as opposed to a large area single crystal cell, makes the reduced performance economically acceptable.

Amorphous semiconductors may be formed as thin films by evaporation or sputtering. Such films have some short-range order due to the directionality of the bonds between atoms, but possess no long-range order. Amorphous semiconductor films differ from polycrystalline films, in that order is maintained over a much shorter distance. Because of the disorder, there will be many dangling bonds, which result in many extra states within the energy gap. For amorphous silicon, it has proved possible to neutralize the effect of the unpaired electrons, by depositing the amorphous silicon in an atmosphere of hydrogen. The hydrogen attaches to the dangling bonds and results in a material, which may be doped p or n type and which can be used to fabricate p–n junctions for solar cells. Although amorphous silicon is relatively cheap to produce, it performs much less well than crystalline silicon.

A1.2.12 Glasses

Glasses are a particular type of amorphous material and are particularly important for optics. They are generally produced from the liquid state by cooling the material. Unlike crystalline materials that have a well-defined melting point, the transition between solid and liquid occurs smoothly over a range of temperatures (the glass transition). They are in essence *super-cooled liquids*, where a large macromolecule is formed as the liquid cools. While most commercially important glasses are very stable over long periods of time, some less stable glasses may revert to a crystalline state more quickly, particularly if held at high temperature or if they are cooled slowly from the molten state. Glasses, unlike crystals, have no long-range order, but because of the preferred directions of the bonds between the constituent atoms, there is an element of short-range order, over a few atomic separations. A consequence of the lack of long-range structure is that there will be density fluctuations that are frozen in as the glass solidifies.

The most common glasses are based on silica (SiO_2) with additions of sodium and calcium oxide (and other metallic oxides) to reduce the temperature at which a glass is formed. Glasses of this type have been produced since prehistoric times and the Romans had a well-developed glass technology. Glasses with special properties are produced by varying the composition. Adding lead oxide increases the refractive index to produce *flint* glass (or lead crystal glass), while the replacement of some of the sodium oxide with boric oxide produces a glass with low thermal expansion (borosilicate glass, e.g. Pyrex). The addition of transition metal oxides, semiconductors or other materials leads to coloured glasses which are frequently used as optical filters. Also, by adding materials such as neodymium or erbium, glasses may be produced which are able to operate as lasers or optical amplifiers.

This ability to control the optical and mechanical properties of glasses accounts for their importance. The development of optical fibres for communications in the 1980s led to the development of new glasses with very low optical attenuation. Such glasses are generally based predominantly on silica, with germania (GeO_2) and B_2O_3 dopants to control the refractive index. In order to achieve

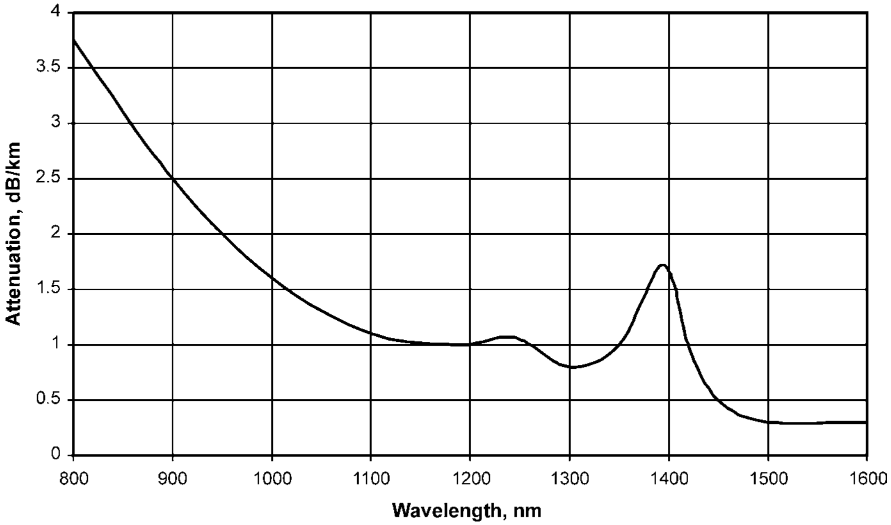


Figure A1.2.21. The attenuation spectrum of a typical silica-based optical fibre.

the very low attenuation required, it is necessary to ensure very high purity, particularly avoiding transition metals. For this reason, the glass is generally fabricated by a vapour phase reaction to generate the basic material, which is usually formed in a molten or finely divided state. The molten glass is then cooled to form the fibre preform rod, which is drawn into a fibre. Using these methods, the attenuation is reduced to a level limited by the scattering from the residual density fluctuations in the glass and the absorption due to the silica. The loss in an optical fibre decreases rapidly with increasing wavelength as the scattering becomes less, until it rises again due to the long wavelength absorption of the glass. Figure A1.2.21 shows an attenuation spectrum for a typical silica-based fibre. The absorption peak around 1400 nm arises from residual hydroxyl radicals in the glass. The success in producing very low loss optical fibres is largely a result of improved fabrication methods, reducing the concentration of these radicals.

While the common forms of glass are all based on oxides, there has recently been much interest in infrared transmitting glasses, for example, chalcogenide glasses that are oxygen free. These glasses are based on the chalcogen group of elements: sulphur, selenium and tellurium, combined with arsenic, antimony or germanium and/or halide elements. Chalcogenide glasses are of interest because of their transparency at longer wavelengths than oxide-based glasses, their semiconducting properties and their optical nonlinearity. In addition, their low phonon energy helps prevent nonradiative decay in infrared optical fibre amplifiers. They are, however, much less robust than conventional glasses softening at much lower temperature (less than 200°C) and showing poor chemical durability.

A1.2.13 Anisotropy and nonlinear optical properties of crystals

A1.2.13.1 Anisotropy

In discussing the refractive index of materials, it was tacitly assumed that the polarizability of a material is a scalar quantity. This must be true for unstrained isotropic materials, such as glasses but is not generally true for crystals. Simple cubic crystals, e.g. sodium chloride, are isotropic, but crystals with more complex crystal structures, e.g. quartz, are not. The relationship between the electric displacement vector, $\boldsymbol{D}$ and the electric field $\boldsymbol{E}$ must be expressed in the more general form:

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{\varepsilon}_\mathrm{r} \cdot \boldsymbol{E} = \varepsilon_0 \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{yx} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{zx} & \varepsilon_{zy} & \varepsilon_{zz} \end{pmatrix} \cdot \boldsymbol{E}. \tag{A1.2.18}$$

It may be shown that the dielectric constant matrix, ε_r, is symmetric and that by suitable choice of major axes, the *principle dielectric axes*, ε_r becomes diagonal. Note that, unless the direction of the electric field, $\boldsymbol{E}$, lies along one of these principle axes the direction of the displacement vector, $\boldsymbol{D}$, and that of the electric field are not in general parallel. The solution of Maxwell's equations for electromagnetic waves in an anistropic medium shows that the propagation depends on the direction and on the polarization state of the wave. For a given direction of propagation, there are generally two normal modes with orthogonal linear polarization, which propagate with different phase velocities. If the wave launched into the crystal is a mixture of the two normal modes, the polarization state will constantly change through a range of elliptical states as the wave propagates.

A1.2.13.2 Electro-optic and nonlinear processes

In discussing the effect of refraction, it was also tacitly assumed that the induced polarization is proportional to the instantaneous electric field. However, for some materials, the interaction between the material and the electric field may be more complex. In the *linear electro-optic* or *Pockels effect*, the interaction is manifest as a change of refractive index with applied electric field. Whether a material shows the linear electro-optic effect or not is determined by the symmetry properties of the crystal. An isotropic material or a crystal with inversion symmetry cannot show the linear effect [10, chapter 6], although it may show the, quadratic, *Kerr effect*. Thus, electro-optic crystals are those having lower symmetry, e.g. quartz or potassium dihydrogen phosphate (KDP). This electro-optic effect finds a use in optical modulators, where electric fields are applied to such a crystal to modulate the polarization or optical phase delay. (A subsequent polarizer or interferometric mixer can convert polarization or phase change to intensity changes.)

If the optical intensity is large, the nonlinearity between the polarization and the applied field may also manifest itself in terms of harmonic generation or optical mixing. The polarization is generally thought of in terms of a power series expansion

$$P = \varepsilon_0 \chi_1 E + \varepsilon_0 \chi_2 E^2 + \varepsilon_0 \chi_3 E^3 + \cdots \tag{A1.2.19}$$

where χ_1 is the linear susceptibility and χ_2 and χ_3 the second- and third-order nonlinear susceptibilities, respectively. The wave equation for propagation of light in a nonlinear medium may be written in the form [10]

$$\nabla^2 \boldsymbol{E} - \mu_0 \mu_\mathrm{r} \varepsilon_0 \varepsilon_\mathrm{r} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = \mu_0 \mu_\mathrm{r} \frac{\partial^2 \boldsymbol{P}_\mathrm{NL}}{\partial t^2} \tag{A1.2.20}$$

where $\boldsymbol{P}_\mathrm{NL}$ contains all the nonlinear polarization terms. The second-order term $\chi_2 E^2$ acts as a driving term, which will contain components at twice the frequency of the electric field of the original wave, leading to the generation of light at a frequency twice that of the initial wave. The second-order component also leads to optical mixing. The third-order term leads to other nonlinear phenomena such as intensity-dependent refractive effects. Nonlinear effects are generally weak, and require high optical intensity and/or a long interaction length.

Many optical devices take advantage of these electro-optic and nonlinear optical effects. For example, the Pockels effect may be used to create an optical phase modulator, or in combination with polarizers, to produce an optical amplitude modulator. Nonlinear optical effects, although usually weak, have long been used to change the wavelength of lasers by frequency doubling or optical mixing. The high optical intensity and long interaction length that is possible in optical fibres have extended the range of nonlinear effects that may be used, without the need for very large optical power.

A1.2.14 Summary

This paper has attempted to provide a brief introduction to the electronic and optical behaviour of some of the materials used. Many of the subjects raised will be taken up again in later chapters and treated in greater depth, notably chapters A1.7 and A2.1. Clearly this is a very extensive field and due to lack of space in this introduction, many issues have been omitted or only treated superficially. One significant area of omission is that of organic and polymeric materials. These materials are becoming increasingly important in a number of areas of optoelectronics, particularly in display technology, which is the basis of a later chapter.

References

[1] Kittel C 1996 *Introduction to Solid State Physics* 7th edn (New York: Wiley)
[2] Myers H P 1990 *Introductory Solid State Physics* (London: Taylor & Francis)
[3] Hook J R and Hall H E 1991 *Solid State Physics* 2nd edn (Chichester: Wiley)
[4] Solymar L and Walsh D 1998 *Electrical Properties of Materials* 6th edn (Oxford: Oxford University Press)
[5] Tanner B K 1995 *Introduction to the Physics of Electrons in Solids* (Cambridge: Cambridge University Press)
[6] Bleaney B I and Bleaney B 1976 *Electricity and Magnetism* 3rd edn (Oxford: Oxford University Press)
[7] Kaye G W C and Laby T H 1978 *Tables of Physical Constants* 14th edn (London: Longman)
[8] Cassey H C, Sell D D and Wecht K W 1975 Concentration dependence of the absorption coefficient for n- and p-type GaAs between 1.3 and 1.6 eV *J. Appl. Phys.* **46** 250–257
[9] Dash W C and Newman R 1955 Intrinsic absorption in single-crystal germanium and silicon at 77 K and 300 K *Phys. Rev.* **99** 1151–1155
[10] Smith S D 1995 *Optoelectronic Devices* (London: Prentice-Hall)

A1.3
Incandescent, discharge and arc lamp sources

David O Wharmby

A1.3.1 Overview of sources

There is a very wide range of incandescent and discharge lamps. The majority of these are sold as general lighting sources, but many are suited to optoelectronic applications. The major lamp companies, and numerous speciality lamp manufacturers also make lamps for applications other than general illumination. Examples of these applications are: projection, video, film, photographic, architectural, entertainment and other special effects, fibre optic illumination including numerous medical and industrial applications, photobiological processes, photochemical processing, microlithography, solar simulation, suntanning, disinfection, ozone generation, office automation, scientific applications, heating etc.

LED sources are covered in detail in chapter B1.1. Section A1.3.10 of this chapter makes some brief comments on the applications in which LEDs are competing with conventional lamps.

This chapter will concentrate on principles and will be illustrated by a number of examples. These principles should make it possible to understand the wealth of information in manufacturer's web sites and catalogues. A selected list of manufacturers is given in appendix.

There are a number of useful books about light sources. The book by Elenbaas [12] is an excellent overview of the science of light sources, whilst for discharge lamps the book by Waymouth [21] contains clear and detailed explanations of many discharge phenomena. Coaton and Marsden [9] give a comprehensive introduction recent enough to cover many modern developments; their appendix 1 gives a useful generic table of lamp data for nearly every commercial source used for illumination. Zukauskas *et al* [24] give an up to date review of the use of LEDs in lighting.

A1.3.2 Light production

Most optical radiation is the result of accelerating electrons and causing them to make *inelastic* collisions with atoms, ions, molecules or the lattice structure of solids. In the UV, visible and near IR, the photons are the result of *electronic* transitions between energy levels of these materials.

There are exceptions; in synchrotron radiation and related processes emission is from accelerated electrons.

As particle densities increase in the source, the spectral features broaden out until, in incandescent sources the spectrum is continuous. Discharge sources generally emit spectral lines of atoms and molecules that are broadened to an extent depending on the pressure. Lamps of various types therefore emit a wide range of spectral features ranging from narrow atomic lines to a full continuum. The types of spectra are often critical for optical applications [5] (see appendix—Oriel Instruments for a selection of spectra).

In incandescent lamps, the radiation is from the surface of a hot material. In discharge lamps, conduction is the result of ionization of the gas; any light emission is a volume process. The task of the lamp designer is to ensure that this ionization is also accompanied by copious radiation of the correct quality for the application.

A1.3.3 Radiation fundamentals

A1.3.3.1 Full radiator radiation and limits on emission

Both in incandescent and discharge lamps, electron motion is randomized. In all cases of practical interest, the drift velocity of the electrons in the applied electric field is much less than the mean velocity. An electron energy distribution function is established that can usually be characterized by an electron temperature T_e. The distribution function may be far from Maxwellian when particle densities are low, or under transient conditions. It is the electrons in the high-energy tail of the distribution that excite the atoms, with subsequent emission of radiation.

The spectral radiance $L_e(\lambda,T)$ of the full radiator or black body is given by Planck's equation (chapter A2.2, where radiometric and photometric quantities are also defined). The spectral radiance is plotted in figure A1.3.1 for temperatures typical of those found in incandescent and discharge lamps. Convenient units for spectral radiance are W m^{-2} sr^{-1} nm^{-1}, obtained by multiplying the value of c_1 in chapter A2.2 by 10^{-9}.

For incandescent or high-pressure (HP) discharge sources the electron temperature T_e is close in value to the temperature T of the solid or vapour, but for low-pressure (LP) discharges in which collisions between electrons and heavy particles are comparatively rare, T_e may be very much higher than the gas temperature. The Planck equation therefore forms a fundamental limit to the radiance that may be obtained from any source in which the electron motion is randomized. This sets a fundamental limit on the spectral distribution, the energy efficiency and the radiance of the source.

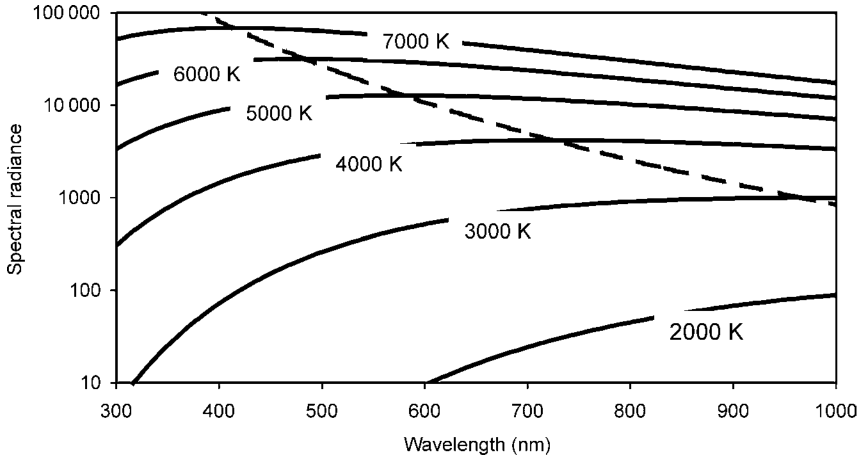


Figure A1.3.1. Spectral radiance of a full radiator (W m^{-2} sr^{-1} nm^{-1}). The broken line is Wien's displacement law showing the shift in peak radiance to shorter wavelengths as the temperature increases.

A1.3.3.2 Absorption and emittance

For radiation falling on a surface

$$a(\lambda, T, \theta) + t(\lambda, T, \theta) + r(\lambda, T, \theta) = 1 \tag{A1.3.1}$$

where the fractions $a(\lambda,T,\theta)$, $t(\lambda,T,\theta)$ and $r(\lambda,T,\theta)$ are known as absorbance, transmittance and reflectance, respectively. In general, they depend on the wavelength, temperature and angle θ between a ray and the normal to the surface.

The spectral emittance $\varepsilon(\lambda,T,\theta)$ is the ratio of the thermal emission from the surface to that of a full radiator (black body) at the same temperature, wavelength and angle. This quantity is also known as spectral emittance. Derived from very general thermodynamic arguments, Kirchhoff's law [13] states that

$$\varepsilon(\lambda, T, \theta) = a(\lambda, T, \theta). \tag{A1.3.2}$$

For a perfect absorber, $a(\lambda, T, \theta) = 1$. Therefore, the spectral emittance of a full radiator is unity; a good approximation can be made by forming a cavity from an absorbing material.

All real materials have $\varepsilon(\lambda, T, \theta) < 1$. The best characterized material is tungsten (figure A1.3.2) [12]. *Selective emittance* is characteristic of most materials; in metals the emittance tails off at long wavelengths, whereas refractory oxides usually have a region of high emittance in the IR.

A1.3.3.3 Étendue

For all optical systems geometry determines how much of the radiation generated by the source can be used by the optical system. This behaviour depends on a very general concept called étendue $\mathcal{E}$, also known as geometric extent [3, 13, 20].

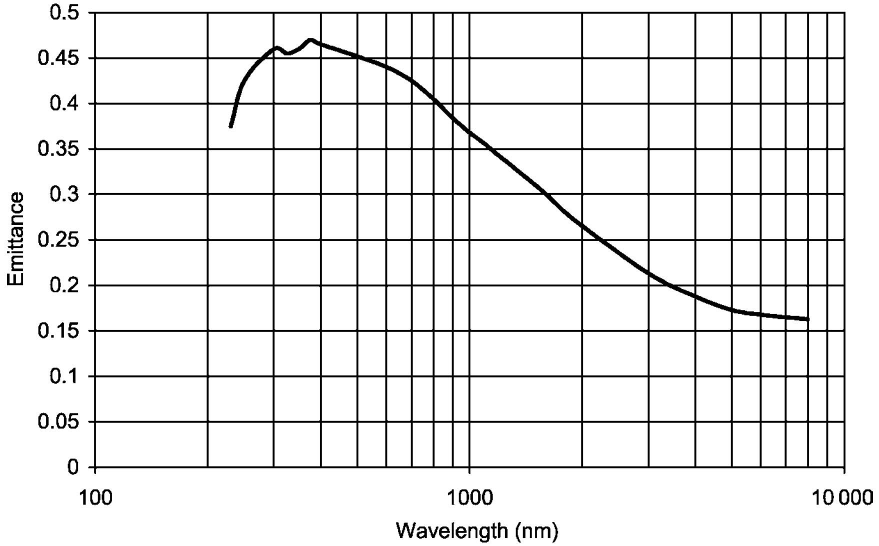


Figure A1.3.2. Spectral emittance of tungsten at 2800 K at normal incidence (after [12]).

A definition of étendue is

$$\mathcal{E} = \iint \cos\theta \, \mathrm{d}A \, \mathrm{d}\Omega \quad (\mathrm{m}^2\,\mathrm{sr}) \tag{A1.3.3}$$

where $\cos\theta\,\mathrm{d}A$ is the projected area of the source under consideration, and $\mathrm{d}\Omega$ is the solid angle into which it is radiating. Notice that the units are geometric, with no mention of amounts of radiation. A more general form is used when refractive indices are > 1 [13]. Energy conservation requires that étendue is conserved in a lossless optical system; if there are losses caused by aberrations, scattering, or diffraction, étendue increases through the system. The étendue of a bundle of rays passing through an optical system either stays the same (ideal) or increases, but never decreases.

A simple example demonstrates some of the issues. Imagine projecting an image of the sun onto a surface. The diameter of the sun is about 1.4×10^9 m with an area $A_\mathrm{S} \approx 1.5 \times 10^{18}\,\mathrm{m}^2$. Our distance from the sun is about 1.5×10^{11} m. Suppose the lens has a focal length of $f = 100$ mm and a diameter of 10 mm so that its area $A_\mathrm{L} \approx 8 \times 10^{-5}\,\mathrm{m}^2$. The solid angle Ω_0 subtended by the lens at the sun is therefore about 3.5×10^{-27} sr. In this simple geometry the étendue $\mathcal{E} = A_\mathrm{S}\Omega_0 \approx 5.4 \times 10^{-9}\,\mathrm{m}^2\,\mathrm{sr}$. The image is brought to a focus at a distance f in a converging beam of solid angle $\Omega_\mathrm{L} = A_\mathrm{L}/f^2 = 8 \times 10^{-3}$ sr. Assuming a perfect optical system so that étendue is conserved, the image area is therefore $A_\mathrm{I} = \mathcal{E}/\Omega_\mathrm{L} \approx 7 \times 10^{-7}\,\mathrm{m}^2$, giving an image diameter of about 0.5 mm.

If we want to focus the sun onto a smaller spot, a lens of the same area needs to have a shorter focal length. Aberrations in a nonideal lens then cause some of the light to fall outside the area predicted above, increasing étendue. Scattering and diffraction are also losses that increase étendue. In general, the integration in equation (A1.3.3) has to be done numerically, e.g. by using an optical design code.

Étendue is also the quantity that determines how the power Φ (W) in the beam is related to the radiance L ($\mathrm{W\,m^{-2}\,sr^{-1}}$), as inspection of the units will confirm:

$$\Phi = L\mathcal{E}t \quad (\mathrm{W}) \tag{A1.3.4}$$

t is the transmittance of the lens (and related optics). Conservation of étendue and of energy means that radiance can never be increased by an optical system.

In a projector, there is always some component that has the smallest (limiting) étendue. Often this will be the film or light gate with its associated projection lens. If the étendue of the source is greater than this, some light will miss the light gate and be wasted. On the other hand, if the étendue of the light gate is much larger than that of the source then the gate will not be fully illuminated. The aim must therefore be to reduce the étendue of the source as far as possible, since it is usually much greater than the limiting étendue. This will minimize the amount of light that misses the light gate. Suppose that a projector lamp has a source of area A_S that radiates in all directions so that the solid angle is 4π and the source étendue is $\mathcal{E}_\mathrm{S} = 4\pi A_\mathrm{S}$. The limiting étendue $\mathcal{E}_\mathrm{L}$ of the system will be usually be that of the light gate. In order that $\mathcal{E}_\mathrm{S}$ does not greatly exceed $\mathcal{E}_\mathrm{L}$, with consequent wastage of light, the area of the source must be very small because the source solid angle is so large. Major advances in projector lamps have been to use HP arcs with an arc gap as small as 1 mm (see section A1.3.7.4) and an effective area in the region of $0.1\,\mathrm{mm}^2$.

The étendue concept is very general. It applies to any illumination system from fibre optics to street lanterns. For example, one of the benefits of LEDs is that their low étendue allows efficient use of the relatively low radiated fluxes; this is a reason why LED headlights for cars are a possibility.

A1.3.3.4 Use of light in systems

The luminous flux in lumens (lm) [9, chapter 1]

$$\Phi_\mathrm{v} = 683 \int_{380}^{780} \Phi_{\mathrm{e}\lambda} V(\lambda)\,\mathrm{d}\lambda \tag{A1.3.5}$$

where $\Phi_{e\lambda}$ is the spectral radiant flux in W nm^{-1} and $V(\lambda)$ is the spectral luminous efficiency for photopic vision (chapter A2.2). The factor 683 (lm W^{-1}) converts power to luminous flux. It is also useful to define the *luminous efficiency of radiation*

$$K = \Phi_v / \Phi_e. \tag{A1.3.6}$$

The (luminous) efficacy of a source is

$$\eta_v = \Phi_v / P_{in} \ \ (\mathrm{lm\,W^{-1}}). \tag{A1.3.7}$$

For many commercial lamps, the input power P_{in} is defined as the power into terminals of the lamp, whereas self-contained sources (such as compact fluorescent lamps), or lamps sold as a system (such as some electrodeless lamps) P_{in} is taken to be the power coming from the electricity supply P_{wall}. The latter power is greater because it contains the losses in the lamp circuit; users should be aware of this possibility for confusion.

Many lighting systems are driven and controlled by electronics and this trend will be maintained in the future. Figure A1.3.3 shows a schematic view of a complete lighting system. To generate light that eventually reaches the eye, every system includes most or all the steps shown. In order to work in terms

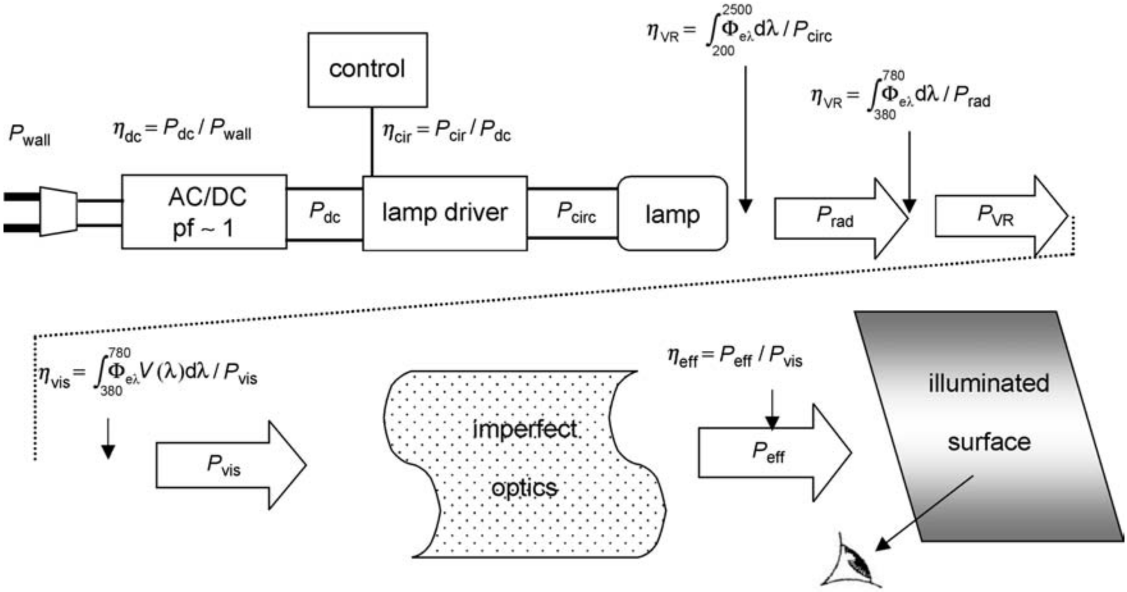

$\eta_{VR} = \int_{200}^{2500} \Phi_{e\lambda} d\lambda / P_{circ}$
control
$\eta_{VR} = \int_{380}^{780} \Phi_{e\lambda} d\lambda / P_{rad}$
P_{wall}
$\eta_{dc} = P_{dc} / P_{wall}$
$\eta_{cir} = P_{cir} / P_{dc}$
AC/DC
pf ~ 1
P_{dc}
lamp driver
P_{circ}
lamp
P_{rad}
P_{VR}
$\eta_{vis} = \int_{380}^{780} \Phi_{e\lambda} V(\lambda) d\lambda / P_{vis}$
$\eta_{eff} = P_{eff} / P_{vis}$
illuminated
surface
imperfect
optics
P_{vis}
P_{eff}

Figure A1.3.3. Schematic diagram of lighting system. The efficiency at each conversion is shown. Power from the wall P_{wall} is converted to dc power P_{dc}, which is used by the lamp circuit to input P_{cir} to the lamp. The broad arrows represent various aspects of the radiation from the lamp. The lamp converts P_{cir} to radiation P_{rad}; a convenient measurement and integration range is 200–2500 nm, which encompasses most of the radiation emitted. A fraction of this P_{VR} is in the visible region. This is then converted into visible power by weighting with the eye sensitivity curve to give power P_{vis}; the luminous flux is $683 \times P_{vis}$. That power is transmitted/reflected by an imperfect optical system onto a surface that reflects light into the eye. All powers P here are in watts and the spectral powers $P(\lambda)$ are in W nm^{-1}.

of power so that we can calculate efficiencies, the quantity in figure A1.3.3 $P_{\mathrm{vis}} = \Phi_{\mathrm{v}}/683$. For each stage in there is a loss and the system efficiency is then

$$\eta_{\mathrm{sys}} = \eta_{\mathrm{dc}} \times \eta_{\mathrm{cir}} \times \eta_{\mathrm{rad}} \times \eta_{\mathrm{VR}} \times \eta_{\mathrm{vis}} \times \eta_{\mathrm{eff}} \tag{A1.3.8}$$

The various terms are defined in figure A1.3.3. Each stage in this chain of light production needs to be examined to discover how system efficiency can be improved. Notice that equation (A1.3.8) applies equally well to a street lamp, a projector, a self-ballasted lamp, a fibre-optic illuminator and, if the conversion from mains to ac is omitted, to battery operated lighting system.

A1.3.3.5 Colour properties and colour temperature of sources

Definitions of quantities mentioned below related to colour are given in chapter A2.2. A comprehensive discussion of colour in lighting is also given by Coaton and Marsden [9, chapter 3].

An important colour property of any source is *colour appearance* or *chromaticity* (specified by the chromaticity coordinates). The colour appearance of any source can be matched by mixture of three sources of different colour appearance (for example, by red, green and blue sources, or by three spectral sources). The space of all possible colours is bounded by the spectral colours. For general illumination and for some opto-electronic applications such as projection, the preferred colour of sources is ‘white’; the chromaticity of these sources is then close to that of a black body having a colour temperature (see below) in the range from about 2800 (yellowish white) to about 8500 K (bluish white). Other sources, such as those used for signalling, usually have more saturated colours (that is colours such as red, green, amber etc.) that are close to the spectral colours). The specifications for these sources are closely controlled [7].

Colour temperature is defined only for those sources having a colour appearance close to that of a black body. The quantity most often used is the *correlated colour temperature* (CCT) [15,19], defined in chapter A2.2. A few examples help to set a scale. The glowing embers of a fire have a CCT in the region of 1000 K whilst a candle flame has a CCT of about 2000 K. Incandescent lamps, depending on type, have CCTs between 2400 and 3400 K. The CCT of the sun is about 6000 K. Discharge lamps for general illumination mostly have CCTs between 3000 and 6500 K. Xenon arcs and flash lamps have CCTs in excess of 6000 K.

Sources of a given chromaticity (that is, having the same colour appearance) may have very different spectral distributions. A commonly observable example (at least in Europe) is that the colour of an amber traffic signal and of the commonly used orange low-pressure sodium street lights are almost identical; the sodium lamp emits only at about 589 nm whereas the traffic signal is a filtered tungsten lamp that emits over a broad spectral range from the yellow through to the red. The value of K (equation (A1.3.6)) for light from the LP sodium lamp also greatly exceeds that for the traffic signal.

Not surprisingly a surface illuminated by these two sources appears to have very different colours. The *colour rendering* capability of a light source is an important measure. For general task illumination colours need to appear ‘natural’; this means that surfaces such as skin, fabric, building materials, etc should not appear distorted when compared with their appearance under natural light or incandescent light, which both have continuous spectra. Along with high efficacy or high luminance, this is a major requirement for commercial light sources. The measure of colour rendering used is the CIE General Colour Rendering Index (CRI) or R_{a} (see chapter A2.2) [6, 23]. R_{a} is computed from the colour shifts shown by a series of coloured surfaces when illuminated by the test illuminant as compared to their colour when illuminated by natural and Planckian reference illuminants.

The better the colour rendering, the lower the efficacy of the lamp. One might think that since the sources that give ‘perfect’ colour rendering have continuous spectra, then high quality lamps should too, and this is usually the case. However, simultaneous optimization of K and R_{a} at constant colour temperature has shown a surprising result; both quantities are maximized if the light is emitted in narrow bands at 450, 540 and 610 nm. This feature of human vision, confirmed by experiment, has been exploited in the *triphosphor* fluorescent lamps that are now standard in all new installations. The lamps use narrow band phosphors that emit close to the critical wavelength. Similar techniques are now being used to optimize white light LEDs (Zukauskas *et al* [24] give a useful review of optimization.)

The CIE CRI is defined so that tungsten and daylight sources have $R_a = 100$. For general lighting in commercial premises requiring high-quality illumination, restaurants and homes, R_a should be 80 or higher. Good quality sources for interior lighting such as triphosphor fluorescent lamps and HP ceramic metal halide (CMH) lamps have $R_a \geq 80$. Lower cost halophosphate fluorescent lamps have R_a around 50–60, as do HP mercury lamps with phosphor coatings. High-pressure sodium (HPS) lamps used for street lighting have R_a around 25.

Human vision is extremely sensitive to small differences in colour [23] particularly in peripheral vision. This has proved to be a major challenge for lamp manufacturers, especially where lamps are used in large installations such as offices and stores. Not only should the initial spread in colour be very small, but also the colour shift during life must be very small otherwise when lamps are replaced it will be very obvious. Amongst the lamps for high-quality illumination triphosphor fluorescent and CMH are pre-eminent in this respect; such colour differences as they have, are barely noticeable.

A1.3.3.6 Radiation from atoms and molecules in extended sources

In a discharge lamp, each elementary volume of plasma emits optical radiation. In a volume source an atom or molecule with an upper state of energy E_u (J) can make a transition to a lower state E_l (J) with a transition probability of A_{ul} (s^{-1}). The emitted wavelength λ (m) is then given by

$$\frac{hc}{\lambda} = E_u - E_l \ \ (\mathrm{J}) \tag{A1.3.9}$$

where h is Planck's constant in J s^{-1} and c is the velocity of light in m s^{-1}.

Since the emission is isotropic, the *emission coefficient* $\varepsilon_\lambda(x)$ from a volume element at position x containing N_u atoms or molecules in the excited state is

$$\varepsilon_\lambda(x) = \frac{10^{-9}}{4\pi} N_u(x) A_{ul} \frac{hc}{\lambda} P(\lambda) \ \ (\mathrm{W\,m^{-3}\,sr^{-1}\,nm^{-1}}). \tag{A1.3.10}$$

$P(\lambda)$ is the line shape function having an area normalized to unity. Do not confuse the emission coefficient ε_λ [18] with the spectral emittance $\varepsilon(\lambda,T,\theta)$ of a surface, which is a dimensionless quantity.

Suppose, we view a nonuniform extended source of depth D. The spectral radiance along a line of sight for a spectral line at wavelength λ is

$$L(\lambda) = 10^{-9} \int_0^D \varepsilon_\lambda(x)\mathrm{d}x \ \ (\mathrm{W\,m^{-2}\,sr^{-1}\,nm^{-1}}). \tag{A1.3.11}$$

This is only an approximation. When absorption is present the radiance does not depend linearly on atom density and is given by the radiation transport equation [18]. Examples of this important phenomenon are described in sections A1.3.6.1 and A1.3.7.1.

A1.3.4 Incandescent lamps

A1.3.4.1 Emission

Tungsten is the pre-eminent material for the manufacture of incandescent lamps. It has a melting point of 3680 K and it can be drawn into the fine wire necessary for making lamps. In normal household bulbs

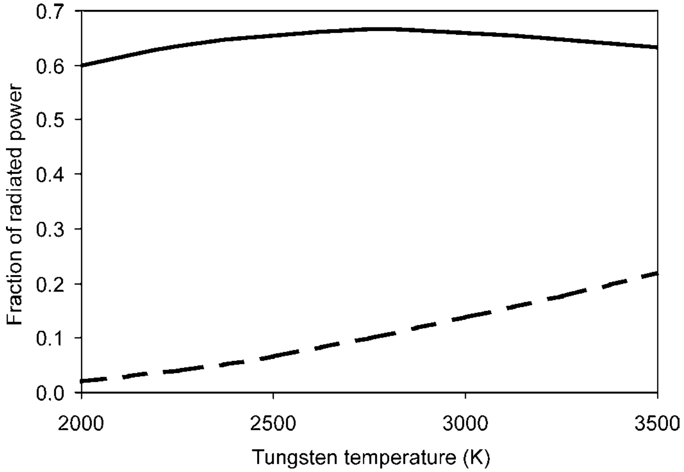


Figure A1.3.4. Calculation of fractions of power emitted from a typical coiled tungsten filament as a function of temperature. The dashed line is the fraction of power radiated at <750 nm (mostly visible but with a very small fraction of UV radiation). The full line is the IR fraction between 750 and 2000 nm. The remaining fraction is at wavelengths >2000 nm (after [1]).

the filament is operated at a temperature in the region of 2800 K, depending on the type. The comparatively low temperature is chosen to limit evaporation and give an acceptable life. This section will concentrate on the higher temperature tungsten–halogen lamps that have many optical applications.

A substantial fraction of radiation from a tungsten filament is emitted between 750 nm and the glass or silica cut-off in the IR. Figure A1.3.4 shows that the fraction of power radiated in the region 750–2000 nm is approximately independent of the tungsten temperature, whilst the visible fraction (<750 nm) doubles for an increase of 500 K in temperature.

A1.3.4.2 Tungsten–halogen lamps

Use of a halogen chemical transport cycle [8] allows tungsten filaments to be operated at higher temperatures than in the standard household bulb. For lamps of similar wattage and life the filament can be operated 100 K higher in a halogen lamp compared with a conventional lamp [1].

The halogen—usually a fraction of a μmol cm^{-3} of iodine or bromine—is added to the lamp before it is sealed. During operation of the lamp the halogen reacts with evaporated tungsten in the cooler regions. The tungsten halide thus produced is a vapour that is transported by diffusion and convection to hotter regions, where it dissociates depositing tungsten and releasing halogen for further clean-up. The dissociation mainly takes place at a region of the filament lead. The net effect of the cycle is therefore to transport the tungsten from the wall to regions of the lamp that do not affect light output.

Because the lamp walls remain clean, the bulb can be made very small and strong. High pressures of inert gas of high molecular weight suppress evaporation. With smaller, stronger bulbs containing high pressures of Kr, or even Xe, the tungsten may be operated at temperatures of up to about 3500 K.

The higher the filament temperature the greater the rate of evaporation and the shorter the life of the lamp. Filaments operating at a colour temperature of 3400 K (filament temperature ≈3330 K) will have a life of a few tens of hours. Life is also strongly dependent on operating voltage; manufacturers' data should be consulted for information.

Tungsten–halogen lamps have the advantage over all other sources of having excellent stability. For best stability, lamps should be operated from a dc constant current supply with current controlled to 1 part in 10^4—this is the technique used for operating calibration lamps. The current should be set to ensure that the voltage rating of the lamp is not exceeded. When lamp stability is at premium (as for example in standards of spectral irradiance) optical equipment suppliers select particularly stable lamps (appendix—Oriel Instruments, Ealing).

A1.3.4.3 Varieties of tungsten–halogen lamps

The development of tungsten–halogen lamps has resulted in thousands of new products being introduced. For optical applications the most important consideration is often the ability to focus the light into a tight beam. This is affected by: the size of the filament, the tightness and evenness of winding of the coil, whether the coil is a flat or cylindrical, whether the coil is concentric with the bulb axis or normal to it, the quality and thickness of the bulb wall, and the type of glass used (hard glass can have better optical quality than fused silica). Examples are shown in figure A1.3.5. High colour temperature versions with powers in the range 25–1000 W or even greater are available. In some cases, these are made from silica that is doped to prevent emission of short wave UV. Consult manufacturers' web sites for 'special' lamps designed for particular optical applications.

A1.3.4.4 Lamps with integral reflectors

There is a wide range of tungsten–halogen lamps built into small reflectors. The reflectors may be aluminized, or have a dichroic (interference filter) coating allowing some IR radiation to escape from the rear of the reflector; this means that the beam is comparatively cool. They may also be fitted with cover glasses that reduce the already small amount of short wave UV that is emitted by fused silica tungsten–halogen lamps. Reflector diameters vary from 50 down to 35 mm. Versions are made with beam

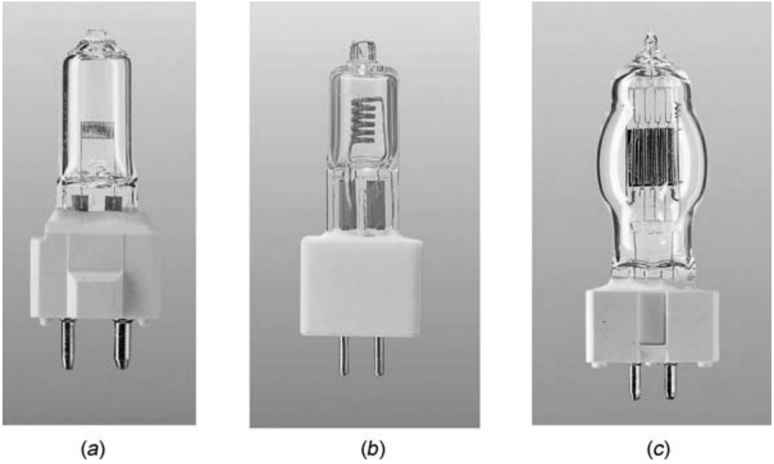

(a) (b) (c)

Figure A1.3.5. Examples of tungsten–halogen lamps (not to scale). All are mounted in a ceramic base that is pre-focused to allow accurate replacement. (*a*) and (*b*) operate from low voltage at a CCT of 3000 K or more and in some cases as high as 3500 K. Lamp (*a*) has a flat filament especially suitable for projection. Lamp (*b*) has an axial filament suitable for use in reflectors for video applications. Lamp (*c*) is a mains voltage lamp for use in overhead projectors; available in ratings up to 900 W and CCT is 3200 K (Philips photographs).

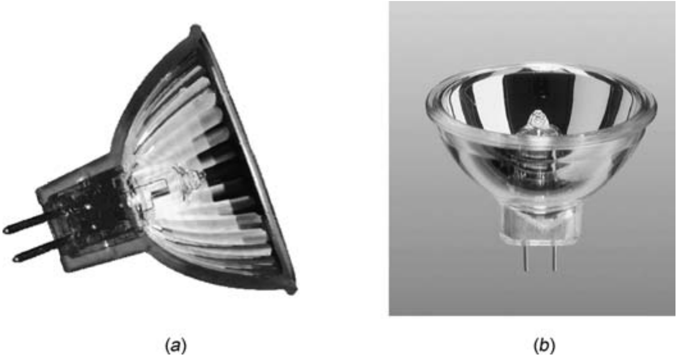


Figure A1.3.6. Examples of low voltage tungsten–halogen lamps in integral pre-focused reflectors. In (*a*) the reflectance of the coating has been reduced to show the positioning of the axial filament in the reflector (Osram). Assembly (*b*) has been specially designed for fibre optic illumination (Philips). Lamps for specific optical purposes are also available from most manufacturers.

divergences from a few degrees up to 40°. Typically wattages vary from 12 to 75 W with a colour temperature of about 3000 K.

These reflector lamps are used in large numbers for all sorts of commercial displays and accent lighting and therefore they are relatively inexpensive. In addition, all the major lamp manufacturers make special versions that are used in a number of optical applications such as overhead projection, microfilm and fibre optic illuminators. Figure A1.3.6 shows examples.

A1.3.4.5 Lamps with IR reflectors

Over 90% of the radiation from tungsten–halogen lamps is in the IR region and so is wasted. Many attempts have been made to return some of this radiation to the filament where it can be absorbed. Commercial success was eventually achieved by using multi-layer interference filters deposited by LP CVD [1]. Less input power is needed to maintain the tungsten coil at the design temperature. The main benefit therefore is a saving in power for a given light output. The beam is cooler since there is less IR radiation emitted although optical quality is degraded slightly by the coating. The main benefit is an improvement of up to 40% in the efficiency of generation of visible light.

A1.3.4.6 IR sources

Incandescent lamps using either tungsten or carbon emitters make use of the IR radiation in industrial heating processes (appendix—Heraeus). The main benefit is a heat source that can be controlled precisely and has a much shorter response time than a conventional oven.

The Nernst source is an example of a ceramic emitter electrically heated to 2000 K, used as an IR illuminator in spectrophotometers. This makes use of selective emittance in the IR. More recent versions of similar devices are given in manufacturers' data (see appendix—Oriel Instruments). There are also low heat capacity carbon emitters that can be modulated at low frequencies (appendix—Hereaus).

A1.3.5 Discharge lamps with electrodes

One way is to group discharge lamps into LP (low-pressure) and HP (high-pressure) types. In LP discharges, the electrons make relatively few collisions per second with the gas atoms and so electron temperature ⩾ gas temperature. In HP discharges, relatively frequent collisions between electrons and gas atoms ensure that both temperatures are approximately equal. The same physical processes occur in LP and HP discharges. Section A1.3.5.1 is concerned with the common features of both types. The electrode regions are described in section A1.3.5.2. Later sections describe their unique features.

Another way to group discharges is by the manner of coupling to the power supply. Most discharges have electrodes in which the cathode is hot; electrons are released into the plasma by thermionic emission. The term *arc* is not uniquely defined, but it is often taken to mean a discharge in which the cathode emits thermionically—examples are all HP discharge lamps and hot cathode fluorescent lamps. In *cold cathode* lamps, the electrodes emit as a result of ion bombardment of the cathode surface. Other discharges (section A1.3.9) are operated at high frequency using induction or microwave sources. Dielectric barrier discharges (DBDs) are transient and self-limiting with little or no emission of electrons from the cathode (section A1.3.9.2).

A1.3.5.1 Stable discharge operation of discharges with electrodes

To start a discharge, a high voltage must be applied to make the gas conducting, and (an electron) current from an external circuit must be passed from cathode to anode through the conducting gas. A by-product of causing the gas to conduct is the production of radiation. To demonstrate the main effects we will consider dc discharges although the majority of commercial lamps operate on ac (section A1.3.8.3).

We will illustrate the main features of a dc discharge using the LP mercury rare-gas discharge of the type used in fluorescent lamps as an example (figure A1.3.7). Other lamps including HP lamps have similar features, but the regions around the electrodes have dimensions that are usually too small to see. The bulk of discharge in figure A1.3.7—the positive column (PC)—is a plasma, so there are equal number of electrons and ions per unit volume. Some discharges such as neon indicators or deuterium lamps used for producing UV are so small that the PC does not exist.

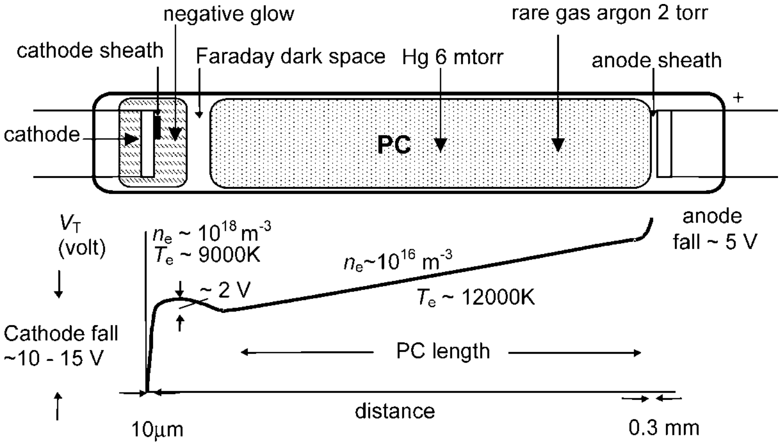


Figure A1.3.7. Structure of a dc discharge. This schematic diagram shows features visible in a typical fluorescent lamp discharge, but they are also present in other discharges. The upper picture shows the positive column (PC), which may be any length, together with the anode and cathode regions in which dimensions are dependent on vapour and pressure. The lower diagram shows the voltage drop V_T along the lamp. The cathode fall field adjusts so that sufficient electrons are extracted to maintain a stable current. Typical electron densities n_e and electron temperatures T_e are shown for the fluorescent lamp case.

In the PC, electrons form a near Maxwellian distribution of energies. Once the discharge has been established, the applied electric field causes the electrons to drift towards the anode and the ions to drift towards the cathode; because their mobility is much greater than that of the ions and the current is carried mainly by the electrons. Therefore, current density is approximately

$$j = n_e|e|\mu_e E \ \ (\mathrm{A\,m^{-2}}) \tag{A1.3.12}$$

where E is the electric field, μ_e the electron mobility, $|e|$ the electron charge and n_e the electron density.

The PC can be any length as long as sufficient open-circuit voltage is available from the supply (think of commercial display signs). A condition for stable operation is that the rate of loss of electrons by recombination with ions must be equal to the rate of gain caused by ionization. In LP discharges, most of the recombination occurs after the carriers have diffused to the wall; in HP discharges, particle densities are high enough for volume recombination to dominate.

The electric field in the column adjusts itself so that electrons are accelerated to a mean energy in the region of 0.5–1.5 eV corresponding to an electron temperature of about T_e of 6000–18000 K. The electron energy distribution then contains enough high-energy electrons to ionize atoms, replacing the electrons lost by recombination. Figure A1.3.7 shows that the electric field in the PC is constant, so in a given gas, the longer the lamp the higher the voltage.

A1.3.5.2 Electrode regions

Adjacent to the anode the voltage usually increases (figure A1.3.7). This is a result of a space charge sheath. If there was no sheath then the anode would only collect the random current. Normally the anode area is too small; to collect the current required it charges positively to attract electrons.

The cathode is more complex [21, chapter 4]. The conditions at the cathode surface have to adjust themselves so that each electron that leaves the cathode initiates events that cause the emission of at least one more electron from the cathode, otherwise the discharge will not be self-sustaining. Electrons emitted thermionically (hot cathode case) or by ion bombardment (cold cathode case) are accelerated in the high field of the cathode fall (CF) region. A beam of electrons from the CF region penetrates the cathode edge of the negative glow (NG) causing the production of positive ions that are accelerated through the cathode sheath. A fraction of these (~0.1) knocks further electrons out of the cathode. The process is entirely self-regulating; if the work function increases, the CF increases and the resulting extra ion bombardment heats the cathode surface to higher temperatures, producing more thermionic emission.

The velocities of electrons leaving the CF are strongly directed toward the anode. This beam is gradually randomized in the direction in the NG region. By the end of the NG they have lost enough energy for the excitation of atomic levels to decrease. This region of comparatively little light is known as Faraday dark space (FDS). At this point, electron motion has been randomized giving a near Maxwellian distribution. Finally as the electrons start to gain energy from the field again, excitation increases and this marks the start of the PC. The NG and FDS therefore serve to change the highly anisotropic electron distribution function coming from the CF into the random distribution in the PC.

In hot cathode lamps, the CF is usually a little greater than the ionization potential of the most easily ionized species (see figure A1.3.7). In cold cathode lamps, the CF is much larger because electrons must be extracted by secondary processes such as ion bombardment. Cold CF voltages are typically in the region of 100–200 V. The CF in cold cathode lamp can be reduced by using hollow cathodes [22].

A1.3.6 Types of LP discharges

By far the most important type is the LP mercury rare-gas discharge used in fluorescent lamps and in UV sources for photochemical and photobiological purposes (section A1.3.6.1). Other LP discharges not described here are LP sodium, used for street lighting of very high luminous efficiency, deuterium lamps used as UV illuminators and LP hollow cathode spectral sources for chemical analysis. There are also a wide variety of LP laser discharges.

A1.3.6.1 Low-pressure mercury rare-gas discharges

LP mercury lamps contain a rare gas, usually argon, krypton or neon or mixtures of these, at a pressure of a few hundred pascal (a few torr). Mercury is added as a small drop of liquid weighing a few milligrams, which collects at the coolest place in the lamp. At typical wall temperatures, the mercury evaporates from the liquid drop at the pressure of about 0.8 Pa (0.6 mTorr). Despite the relatively low number density of the mercury atoms they dominate the properties of the discharge. The fluorescent lamp discharge is a highly efficient emitter of UV in the mercury resonance lines at 254 and 185 nm (>70%).

Phosphors are used to convert UV to visible radiation [9, chapter 7]. Lamp phosphors are ionic materials doped with activators that absorb at short wavelengths and then re-emit at longer wavelengths. The energy deficit in this *Stokes*' shift is converted into lattice vibrations. In fluorescent lamps used in lighting the conversion loss is typically 50%. There is a very large range of phosphors [9, chapter 7], and fluorescent lamps giving white light of many different CCTs and other colour properties are available. Particularly important are the ionic rare-earth based phosphors as these emit at the wavelengths that combine high efficacy and colour rendering index (R_a—section A1.3.3.5). The principle ones are noted in table A1.3.1; notice how close the peaks are to the 450, 540 and 610 nm wavelengths that optimize colour rendering and efficacy.

The notation in table A1.3.1 is the chemical composition of host lattice:activator. The activator is an ion added deliberately at relatively small concentrations to absorb UV and emit visible light. In some cases, the host lattice has this same function. Quantum efficiencies are close to unity.

One very important benefit of rare-earth phosphors is their resistance to degradation by mercury discharges at high power leadings. It is this property that made possible the development of compact fluorescent lamps. The disadvantage is the high cost of rare-earth phosphors compared with the halophosphates that they have largely replaced. The complexity of the materials in table A1.3.1 is such that phosphor research is still largely empirical, so the existence of each of these phosphors represent many man-years of painstaking research.

Table A1.3.1. Phosphors commonly used in fluorescent lamps.

Name	Formula	Wavelength of peak output (nm)
YEO	$Y_2O_3{:}Eu^{3+}$	611
CAT	$Ce_{0.65}Tb_{0.35}MgAl_{11}O_{19}$	543
LAP	$LaPO_4{:}Ce^{3+},Tb^{3+}$	544
CBT	$GdMgB_5O_{10}{:}Eu^{2+}$	545
BAM	$BaMg_2Al_{16}O_{27}{:}Eu^{2+}$	450
Halophosphate	$Ca_5(PO_4)_3(F,Cl){:}Sb^{3+},Mn^{2+}$	broad bands

The mercury vapour pressure is a dominant factor in controlling the amount of radiation emitted and the efficiency with which it is generated. When a mercury atom is excited near the centre of lamp, the emitted photon is at exactly the correct energy to be absorbed by a ground state atom nearby. The photon is absorbed and reabsorbed many times before it finally reaches the wall in a random walk. When the mercury pressure is high, there are so many steps in the random walk that the chance of losing the excitation energy nonradiatively in a collision increases. When the mercury vapour pressure is low the initial excitation energy can escape in a small number of steps, but then the fraction of collisions that lead to excited mercury atoms is low. (The related process in HP lamps is described in section A1.3.7.1.)

This means there is a mercury vapour pressure at which the efficiency of generation of UV radiation is at a maximum. This optimum pressure is achieved by having a small amount of liquid mercury present at about 42 °C. When using fluorescent lamps it is important to arrange for the fixture or unit holding the lamps to operate so that the mercury pressure is close to optimum. Lamps are designed to run close to optimum in commercial lighting fixtures. For other uses, such as backlighting some cooling may be necessary. Some types of multi-limb compact fluorescent lamps are designed for operation in hot fixtures. In these, the mercury is dosed as a solid amalgam containing, for example, bismuth and indium. The vapour pressure of mercury above the amalgam is less than that above free mercury, but the use of an amalgam also substantially increases the ambient temperature range over which the mercury pressure is close to optimum [2].

A1.3.6.2 Applications of LP mercury discharges

The fluorescent lamp discharge lends itself to many different formats [9, chapter 7]. The most familiar are the long thin lamps used in ceiling lighting in nearly all commercial and industrial premises. There are also a wide variety of compact fluorescent (CFL) designed as a high efficiency replacement for incandescent lighting.

Other than illumination, important applications for fluorescent lamps are in office equipment (copiers, fax machines, etc) and in the backlighting of displays. Cold cathode fluorescent lamps have a number of benefits: they can be small in diameter allowing screens to be very thin; at the low powers needed they are efficient enough for the purpose; lives are long; they can be switched frequently; and low cost, efficient power supplies are readily incorporated in the end product. Hot cathode fluorescent lamps produce more light and can be used to backlight displays that are used in high ambient light levels such as ATM machines. Short wave radiation from hot cathode mercury rare-gas discharges is used in photochemical or photobiological processes; or it can be converted using a phosphor to UVA (as in 'black light' sources) that show up fluorescence in materials.

A1.3.7 HP discharges

There are many variants of HP discharges. Most of them are used for street lighting and interior illumination of stores and offices and other commercial premises, in which high luminous flux, high efficacy, good colour quality and long life are at a premium. Lamps exist in *single-ended* (both connections at one end) and *double-ended* (one connection at each end) configurations to suit different applications. Many other types of HP discharges are used in which light must be projected and high brightness is needed. Some of the properties of HP discharges are described below. The two main classes of lamp are those that use volatile or gaseous elements, and those that use metal halides to introduce radiating species into the vapour.

A1.3.7.1 General features of HP discharge lamps

We will illustrate the operation of HP discharges by using the HPS (high-pressure sodium) lamp as an example. An HPS lamp has electrodes inserted into a narrow arc tube made from translucent alumina, resistant to attack from sodium. As with many HP lamps the arc tube is contained within a glass outer bulb. These lamps are used as highly-efficient (120 lm/lamp watt) long-lived ($>$20000 h) street lights that give a pleasant golden light with CCT = 2000 K, albeit with rather poor colour rendering properties ($R_a = 25$).

The dimensions of the tube are typically 7 mm internal diameter with 70 mm length between the electrode tips for 400 W rating, with dimensions decreasing for lower wattage lamps. They contain a small pressure of rare-gas and a few milligram of sodium metal. On applying a voltage the rare gas breaks down. The resulting discharge heats and evaporates sodium until its pressure is about 1.4×10^4 Pa (100 Torr). A radial temperature profile develops in which the centre temperature is about 4000 K and the wall temperature is about 1500 K. Most of the length of the discharge is a positive column uniform along the axial direction. Sodium lamps usually also contain about 10^5 Pa (760 Torr) of mercury vapour. This reduces thermal conduction and increases axis temperature, thus increasing spectral radiance.

The positive column is approximately in local thermodynamic equilibrium (LTE) [18]. This means that the properties are dependent on the local temperature in the plasma. The electron density is given by a version of the law of mass action called the Saha equation [10]

$$\frac{n_e n_i}{n_a} = S(T) \quad (\mathrm{m}^{-3}) \tag{A1.3.13}$$

where n_e and n_i are the electron and ion densities, n_a is the density of atoms (number per m^3) and

$$S(T) = 4.83 \times 10^{21}\,(U_i/U_a)\,T^{3/2}\exp(-E_i/kT) \quad (\mathrm{m}^{-3}) \tag{A1.3.14}$$

where the U factors are partition functions for the ion and atom. $S(T)$ depends strongly on temperature through the exponential factor, where E_i (J) is the ionization potential (including corrections for high electron density) and k is Boltzmann's constant. Since the hot gas is a plasma $n_e = n_i$. The atom density n_a in an elementary volume at temperature T is given by the gas law so $n_a = P/kT$ where P is the gas pressure. Table A1.3.2 shows values of n_e in sodium vapour at various temperatures. Since the current density is proportional to n_e it is clear that the current flow is mainly in the high temperature region.

The population n_u of an energy level of an atom (labeled u) is given by another LTE formula:

$$n_u = \frac{g_u}{g} n_0 \exp\left(\frac{-E_u}{kT}\right) \quad (\mathrm{m}^{-3}) \tag{A1.3.15}$$

where n_0 is the density of atoms in the ground state, E_u is the energy (J) of the upper state of the atom, whilst g_0 and g_u are the statistical weights of ground and upper states, respectively. The number of atoms excited to the upper state depends exponentially on temperature. Because of the exponential *Boltzmann factor* in equation (A1.3.15), the fraction of atoms in the excited state u is very small even at the highest temperatures. Only in the hottest parts of the discharge are significant numbers of atoms excited; the resulting 'corded' appearance is a characteristic feature of an LTE arc. When a HP discharge operates horizontally convection bows the bright part upwards—the origin of the term arc. The importance of equations (A1.3.13) and (A1.3.15) is shown in table A1.3.2.

Self-absorption dominates the spectrum of many HP discharge lamps and is especially dominant in HP sodium discharges. As figure A1.3.8 shows there is no significant radiation at 589 nm, the wavelength at which sodium radiates at low pressures. In an HPS lamp, the sodium pressure is so high that photons

Table A1.3.2. Shows how the plasma temperature affects the number density (m^{-3}) of excited states and ions. Electron density is equal to ion density. There are two excited states at about 2.1 eV giving rise to the characteristic orange sodium D radiation. The ionization potential is 5.14 eV before correction is made for lowering of the value at high electron densities. The arc operates so that the electron density is sufficient to carry the current and the plasma temperature adjusts to make this so. For steady state sodium arcs this sets the maximum plasma temperature to about 4000 K. Calculated using equations (A1.3.13) and (A1.3.15).

	Plasma temperature (K)			
	2000	3000	4000	5000
Number density of sodium atoms	4.8×10^{23}	3.2×10^{23}	2.4×10^{23}	1.9×10^{23}
Fraction of sodium atoms excited to the states radiating at 589 nm	1.5×10^{-5}	8.8×10^{-4}	6.4×10^{-3}	2.3×10^{-2}
Fraction of sodium atoms that are ionized	7.3×10^{-6}	1.9×10^{-3}	3.5×10^{-2}	2.2×10^{-1}

from excited sodium atoms can only travel about 10^{-7} m at the line centre before being absorbed by a ground state atom. However, there is a chance that very close collisions with other sodium atoms can perturb the radiating atom sufficiently so that it radiates at wavelengths far from the line centre at 589 nm. The hot plasma can therefore be considered as storing excitation energy until the energy can escape from an atom having strongly perturbed energy levels. The higher the pressure, the further from the line centre the wavelength has to be, before the light can escape (figure A1.3.8). This behaviour is called self-reversal and it has a dominating effect on the operation of many HP discharges [9, section 5.6.3]. The cover of the book by de Groot and van Vliet [10] shows beautiful colour photographs of the self-reversal of the sodium D lines at different pressures.

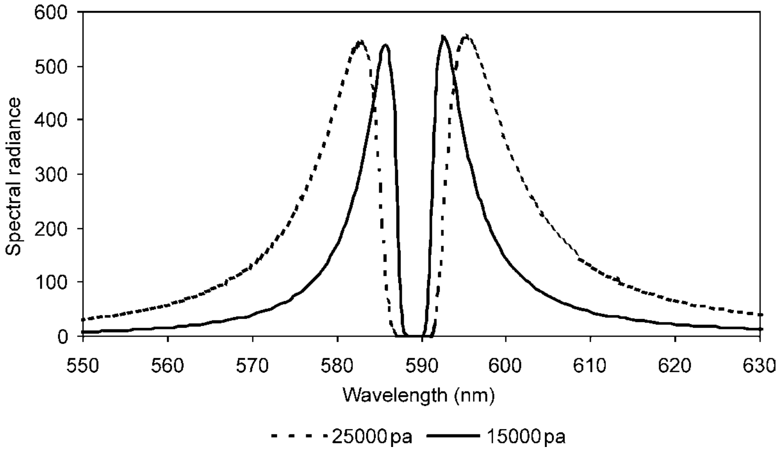


Figure A1.3.8. The formation of self-reversed lines in high-pressure sodium lamps at two sodium pressures. The calculation has been done for a parabolic radial temperature profile for a centre temperature of 4000 K and a wall temperature of 1500 K. Comparison with figure A1.3.1 shows that the peak radiance is substantially lower than that for a black body at the maximum temperature.

A1.3.7.2 HP metal halide lamps

There are very few elements that have well-placed spectral lines *and* sufficiently high vapour pressures to be operated as HP discharges, the most important being mercury, sodium, sulphur, and the permanent gases (of which Xe is by far the most important).

There are perhaps 50 elements that have metal halides that are sufficiently volatile to be used in HP lamps. The principal ones are as follows:

- Na, I, In, Tl, Ga halides in which the metals have relatively few strong atomic lines
- Sc, Fe, Dy (and other rare earth) halides in which the metals have many relatively weak visible lines so close together that the spectrum appears continuous at moderate resolution.
- Sn, Pb and similar halides that form relatively stable monohalide molecules that emit a spectrum that appears continuous at low resolution.

The halide is usually the iodide, which has the least reactive chemistry.

Metal halide lamps are extensively used as efficient white light sources of good colour quality for general illumination; all major lamp manufacturers make them (appendix). Because the spectrum can be tailored to use, metal halide lamps are used extensively for production of UV for photo-polymerization processes, such as ink drying or glue curing (appendix—Heraeus). These lamps are installed as part of large production processes in the printing and packaging industries. The speed of curing is often the bottleneck in the processes, so with suitable UV sources productivity can be increased. HP mercury lamps are also used for similar processes. Other uses of metal halide lamps include special versions for medical conditions such as psoriasis (appendix—Osram).

A1.3.7.3 Operating principles of metal halide lamps

Many metal halide lamps contain thallium iodide (TlI). TlI is considered as a simple case to illustrate how the light is produced in metal halide discharges. Figure A1.3.9 shows a schematic diagram of a HP TlI discharge. When the lamp is made, a few milligram of solid TlI and a rare gas for starting are added. Usually enough mercury is added to give a partial pressure of about 10^6 Pa (10 bar) to reduce thermal conduction and to adjust operating voltage (section A1.3.7.1). When the lamp is operated, the rare gas discharge heats up the Hg and TlI causing them to evaporate. In higher temperature regions, the TlI dissociates into Tl and I atoms. At higher temperatures still the Tl is excited and emits intense green light of high efficiency that can be useful for underwater illumination. Finally, near the axis the Tl is ionized producing the electrons needed to carry the current. This progressive evaporation, dissociation, excitation and ionization occurs in all metal halide discharges.

With mixtures of halides the ratio of salts has to be chosen with due consideration to the chemistry of the liquids and vapours. For example, one of the first types of metal halide lamp used mixtures of indium, thallium and sodium iodides that emit blue, green and orange self-reversed spectral lines. Altering the proportions of these can provide white light discharges of different colour temperatures and quite good efficacy (luminous efficiency). However, their colour rendition is rather poor.

Metal halide arc tubes are generally shorter than HPS lamps (for which the length to diameter ratio is more constrained by requirements of optimization) and may even be close to spherical in shape. This has an effect on étendue and may make fixtures using these lamps more efficient at using the light.

There has been extensive research and development over the last 40 years that has produced mixed metal halide lamps with much improved colour performance and efficacy. The halides used, their vapour

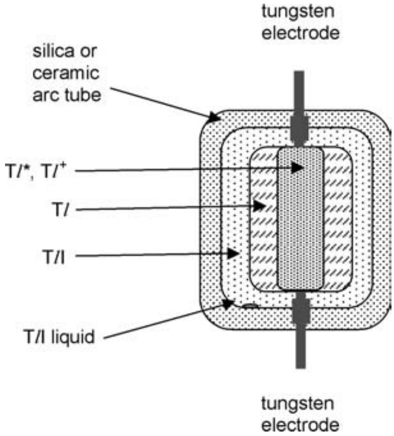


Figure A1.3.9. The principle of operation of a metal halide discharge. In this example solid thallium iodide (T*l*I) is dosed into the lamp along with a rare-gas. The discharge starts in the rare gas, melting and then vaporizing the T*l*I. In the steady state the current is provided by ionising T*l* atoms. For this to happen the temperature on axis needs to be above 5000 K. At this temperature T*l* atoms radiate strongly with their characteristic green line at 535 nm. The boundaries between the various regions are not sharp as shown schematically here, but blend into each other as the temperature increases from the wall to the axis. In a practical lamp, mercury vapour is also introduced at several bars to reduce thermal conduction losses.

pressures and their relative proportions all have a strong influence on the initial colour properties and efficacy.

It is important that these properties stay constant through lives of 10 000 h or more. Reactions between the various components and the tube walls occur at different rates; all metal halide lamps show some colour shift during life. Detailed R&D has improved the colour stability of metal halide lamps in silica arc tubes so that it is acceptable in critical applications such as the lighting of stores and offices. A recent major improvement has been in the use of translucent alumina ceramic arc tubes for containing Na, Dy, Tl, HgI metal halide arcs. Metal halide reactions with the envelope are much slower than with silica and this has provided a further major improvement in initial colour uniformity and colour stability through life.

Most metal halide lamps are used for illumination where the transparency of the arc tube is not an issue, but the scattering by the arc tube is a major disadvantage for projection. For projector and automotive head lamps, silica arc tubes are universally used. Because the axis temperature of metal halides is usually around 5500 K, the radiance of the gas close to the axis can be very high (see figure A1.3.1).

A1.3.7.4 Applications of HP discharge lamps

Table A1.3.3 gives information about HP lamps used for general illumination. It is intended to show the range of types available with some idea of the best characteristics to be expected. All types exist in more than one power rating, but the rating given is a fairly typical one for the indicated application and type. Generally efficacy increases as power rating increases [9, appendix 1].

Table A1.3.3. Indicative characteristics of HP discharge lamps used for general lighting.

Lamp type	Application example	Power (W)	Initial (lm W^{-1})	Life (10^3 h)	CCT (K)	R_a
HPS	Road lighting	400	125	30	2000	25
High CRI	Prestige town lighting	400	100	24	2200	60
HP mercury vapour + phosphor	Road lighting	400	60	24	3500	55
Metal halide	Prestige outdoor, stores	400	90	24	4000	70
CMH	Commercial interiors	100	90	12.5	3000	85

Most of the lamps that are in the table are arcs with positive columns of length of several centimetres that are stabilized by the tube wall. The CMH lamp is a short arc lamp in which the arc is mainly stabilized by the electrodes.

Manufacturers' web sites give many examples of applications other than for illumination. All the major lamps manufacturers make a variety of metal-halide and xenon short arc lamps for projection and related uses (appendix). In short arc lamps (length $<$ few millimetres) there are usually regions close to the electrodes that have particularly high arc temperature. This region of high arc temperature forms because the electrodes cool the arc, and the field close to the electrodes has to increase to maintain conduction; moreover the current density normally increases as the arc contracts toward the cathode hot spot. The combined increase in current density and field means that the power per unit volume of arc is greatest just adjacent to the electrode. Although this generally leads to a reduction of efficacy there may be an increase in luminance. Figure A1.3.10 shows examples of lamps that are used for a variety of projection and entertainment applications and other more specialized applications such as solar simulators.

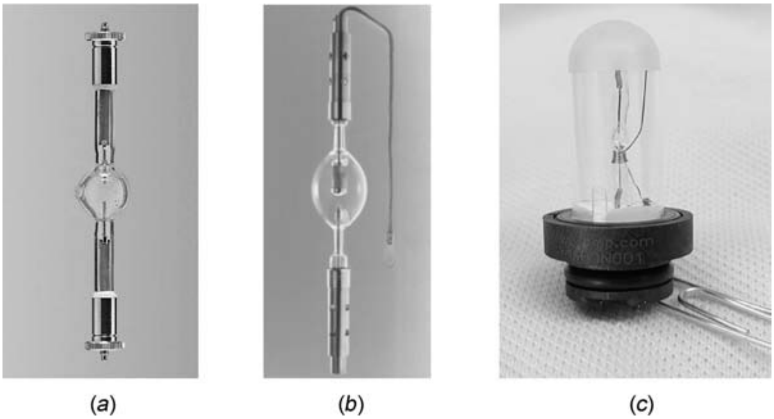


Figure A1.3.10. The large range of powers possible with HP discharges is shown. Not to scale. (*a*) An example of a metal halide lamp for entertainment applications with a CCT of about 6000 K, available in ratings between 575 and 12000 W (Philips). (*b*) An example of a high pressure xenon lamp operated from dc with a CCT of 6000 K, available in ratings between 450 and 12000 W (Osram). (*c*) A 10 W metal halide lamp operating with CCT around 6000 K and ratings between 10 W (Welch-Allyn).

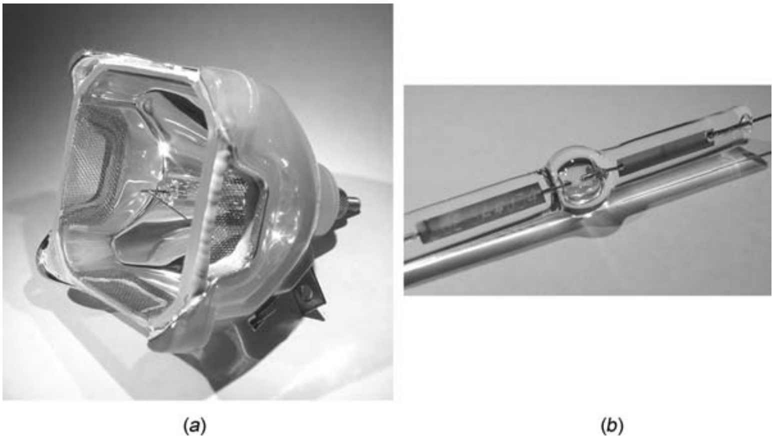

(a) (b)

Figure A1.3.11. (*a*) A pre-focused projection unit designed for LCD projectors. The reflector is carefully designed to keep étendue of the pre-focused unit as low as possible and has a dichroic filter to reduce the amount of IR in the beam. (*b*) The arc tube used in the reflector. It operates at 130 W and a CCT of about 6200 K with a mercury pressure of more than 150 bar. The arc gap of about 1.2 mm with consequently low étendue. (Philips, Osram, GE Lighting.)

Short arc metal-halide lamps can readily be integrated into pre-focused parabolic or elliptical reflectors (appendix—Welch Allyn, Ushio and others). Various beam divergences are available to suit different applications. A relatively recent development is shown in figure A1.3.11 (appendix—Philips, Osram and GE Lighting). This very high-pressure (1.5×10^7 Pa or 150 bar) has extremely high luminance because of its extremely high arc temperature. The reason for the high arc temperature is that 130 W are dissipated in an arc of length hardly more than a millimetre. Spectral lines show extreme broadening and there is an intense continuum giving good colour rendition. With an arc gap of only 1.2 mm the étendue is very small. The lamps are designed to be operated in a pre-focused reflector and the whole assembly used in data or video projectors.

A1.3.8 Electrical characteristics of discharges

A1.3.8.1 Breakdown and starting in discharge lamps

The gas in the lamp must be converted from an excellent insulator into a good conductor with a resistance that can be as low as a few ohms. Figure A1.3.12 shows the voltage across the lamp as a function of current over a very wide range of currents. After breakdown, the current increases rapidly until finally it stabilizes at the value needed to satisfy the circuit equations. In order to start the lamp, the circuit must be able to provide a voltage in excess of the highest lamp voltage in this diagram.

In order to achieve breakdown some source of electrons is necessary. If not provided by other means, they result from ionization by cosmic rays or natural radioactivity in the materials of the lamp. In other cases reliable breakdown is aided by the addition of small amounts of radioactive materials such as Kr^{85}, or by photoemission from surfaces caused by a small external source of UV. In hot cathode fluorescent lamps the electrodes can be heated before the voltage is applied: at low temperatures the field-enhanced thermionic emission provides enough electrons [21]. In HP lamps, a third trigger electrode is often included adjacent to the main electrode. When the voltage is applied across this small gap, breakdown is

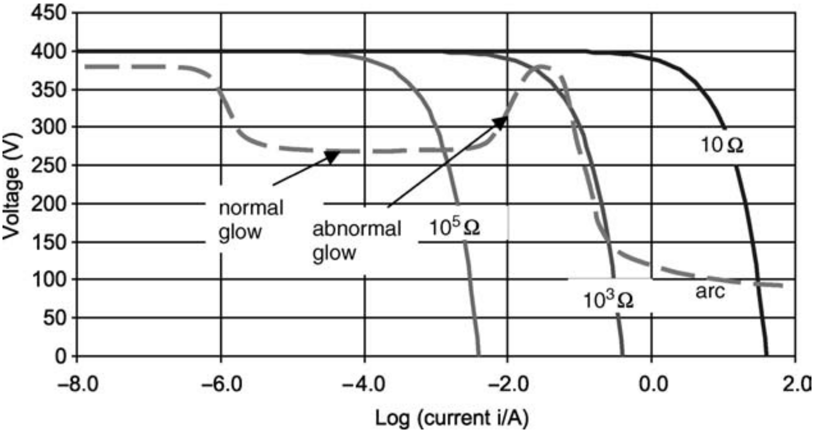


Figure A1.3.12. Discharge voltage as a function of current (dashed line) over a wide range of currents. The well-known discharge regions are shown. The voltage in the arc region is small because of the low cathode fall resulting from hot cathode operation. The load lines for various values of series resistance are shown with intersections in the normal glow, abnormal glow and arc regions. The intersection in the abnormal glow region may not be stable.

assured; this gap then provides initial electrons for the main gap. A general rule is that the fewer the initial electrons the higher the starting voltage needs to be, and the longer the time lag before breakdown.

The majority of lamps operate with hot cathodes that emit thermionically. Once breakdown has been achieved the transition from abnormal glow to arc in figure A1.3.12 must be achieved quickly and cleanly. Staying too long in the region of cold cathode operation, in which the electrons released by ion bombardment can be very damaging and shortens lamp life, sometimes dramatically.

There are various schemes for starting fluorescent lamps [21] in which the electrodes are tungsten coils coated with electron emission mix. One of the most common is to use a starter switch in parallel with the lamp. The starter switch is wired so that when closed, current limited by the ballast is passed through both electrodes. Initially the starter switch is closed. This preheating process raises the temperature of the electrode to about 1000 K before the starter switch opens. On opening the switch the open circuit voltage plus the self-induced voltage across the inductance is applied across the lamp, causing breakdown. The main purpose of preheat ensures that thermionic emission occurs very soon after breakdown. The switch is usually a relatively inexpensive bi-metallic type. The tolerances on starter switch operation are closely constrained according to the lamp type. Increasingly fluorescent lamps are operated from electronic ballasts. It is a relatively simple matter to include a precision electronic preheat circuit to enhance the lamp life. At the low cost end, preheat is not included; after breakdown the electrode is heated rapidly by ion bombardent until it reaches thermionic emitting temperatures. Up to this time secondary emission dominates but the ion bombardment heats the cathode; these so-called instant start lamps generally have shorter life than lamps that are preheated.

For HP lamps, preheating is not an option. Starting from cold when the pressure is around 10^4 Pa (0.1 bar), breakdown voltages in the region of some kilovolts are needed. After an HP lamp has stabilized the pressure may be many atmospheres; on turning off it will require some tens of kV to restart immediately. Various types of pulse ignitor are used. The speed of transition from the glow to the arc is sensitive to many factors related to electrode design, lamp fill, processing quality and the open circuit voltage available. Ballast and lamp designers work together to ensure that the glow to arc transition in figure A1.3.12 occurs rapidly and cleanly to ensure long life.

A1.3.8.2 Steady state electrical characteristics

For an *ohmic* conductor the number of carriers is independent of current, so changing the voltage simply changes the mean drift velocity of the carriers. As long as the temperature remains constant the current is proportional to voltage—Ohm's law. If the temperature increases the carrier mobility decreases and the resistance increases. This is what happens in tungsten lamps in which the hot resistance can be 15 or more times higher than the cold resistance.

Discharges show strongly *nonohmic* behaviour (figure A1.3.12). A discharge is a current-controlled device; and the voltage between the terminals sets itself to maintain this current. An additional impedance called a *ballast* is necessary to control the current. In response to increasing current hot cathode discharges respond by decreasing the voltage across their terminals. This is the so-called negative, or falling V–I characteristic. The rate of decrease of voltage with current is usually quite small. This corresponds to the arc region on the right hand end of figure A1.3.12 where the lamp voltage is comparatively low.

Figure A1.3.12 shows what happens to voltage as the current is increased over many orders of magnitude. Increasing from low values of current the voltage decreases to a plateau region in which a glow is visible on the cathode. On increasing the current, the glow increases in area whilst the voltage remains constant, implying that the current density at the surface of the cathode is constant. This is called the *normal glow* regime. As the current is increased further, the glow finally covers all the cathode area and often the leads as well. At this point, the current density has to increase and the voltage across the terminals increases. This is called the *abnormal glow* region. In both the normal and abnormal regions, the major part of the lamp voltage is dropped across the CF. The resulting ion bombardment increases the cathode temperature and the cathode begins to emit thermionically and makes a transition to the arc regime, which has a low CF. The abnormal glow region has a positive resistance characteristic, but this is not stable unless there is a ballast in the circuit.

For a dc discharge a series resistance R is needed to stabilize the current I. If the supply voltage is V_S and the lamp voltage V_T then

$$V_T = V_S - IR \quad (\mathrm{V}) \tag{A1.3.16}$$

The right hand side of this equation is called the load line. Load lines for three resistances are shown in figure A1.3.12. For the highest resistance the intersection point is in the normal glow region (typical of a neon indicator lamp). With the lowest resistance the intersection is in the arc region. The intermediate resistance has two intersections, the one at the lowest current is in the abnormal glow region. If the heating of the cathode is insufficient to cause a transition to an arc, then the lamp remains in the abnormal glow condition. In some cases when starting an arc the discharge sticks in the abnormal glow with a high cathode fall; the sputtering can then cause very rapid blackening of the walls and premature failure.

Despite what the manufacturers' data sheets may say, figure A1.3.12 suggests that there is no specific power at which a discharge must operate; adjusting lamp current by using ballast impedance and supply voltage means lamps may be operated at a wide range of powers—at least for a time. The consequences of operating at powers different from the rated power are usually a reduction in life; properties such as colour temperature and colour rendering and efficacy will also change. Nevertheless, for specific applications this is an option that the user can consider.

The reason for needing a ballast is best explained by using an argument given by Waymouth [21, chapter 2]. Figure A1.3.13(*a*) shows the falling arc characteristic (part of the right hand end of figure A1.3.12). This characteristic is the locus of points for which $\mathrm{d}n_e/\mathrm{d}t = 0$. The further above this line the more the rate of ionization exceeds the loss, so the current increases, and this increases $\mathrm{d}n_e/\mathrm{d}t$, with the result that the current continues to increase. If the applied voltage is below the line the loss exceeds production, the current decreases and the discharge extinguishes. Figure A1.3.13(*b*) shows the effect of a

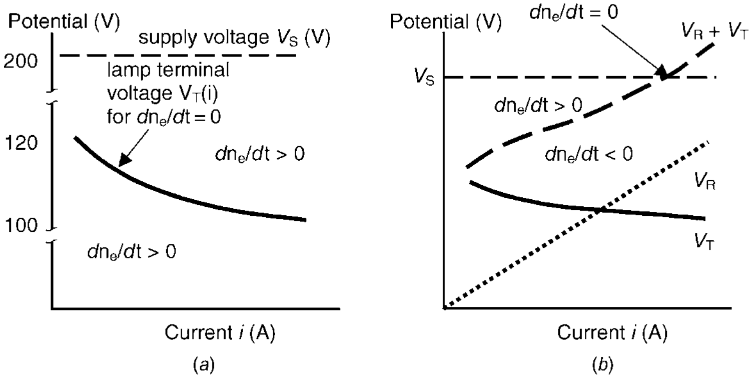


Figure A1.3.13. (*a*) A section of the arc region of figure A1.3.9. The solid curve is the locus of points for which $dn_e/dt = 0$. If a discharge without a ballast could be prepared at a point on this line small fluctuations would cause the current to increase without limit or decrease to zero. (*b*) Shows the effect of a stabilizing ballast resistor. The voltage across the circuit is now the sum of the lamp voltage V_T and the resistor voltage V_R. If the current fluctuation increases the current, the sum of resistor and lamp voltage increases into a region where $dn_e/dt < 0$ and the current immediately decreases again until $V_R + V_T = V_S$. This operating point is therefore stable against lamp current fluctuations.

series ballast resistor. The total circuit voltage $V_T + IR$ now intersects the supply voltage at a certain current. If a fluctuation causes the current to increase, then the total circuit voltage moves into a region where $dn_e/dt < 0$, thus immediately decreasing the current. If the current decreases, then the total circuit voltage decreases so that $dn_e/dt > 0$, thus increasing the current again.

A1.3.8.3 AC operation

Resistive ballasts work satisfactorily, but are lossy. Commercial lamps operate from the ac mains supply using magnetic inductances as ballasts [9, chapter 17]. Figure A1.3.14 shows the lamp voltage and current waveforms for a fluorescent lamp on a resistive ballast. At 50 Hz there is an appreciable re-striking voltage after current zero. This extra voltage is needed to restore the electron density after it has decayed during the latter part of the previous cycle. If this re-strike voltage exceeds the supply voltage the lamp will extinguish. The phase relationships in an inductive circuit mean that a large voltage is available at the time that the current reverses, so extinction is less likely. For stability on ac mains supplies with a series inductance, the rms lamp voltage should not exceed about half of the rms mains voltage.

Most lamps are now developed to operate from electronic power supplies. Although more expensive than magnetic ballasts, there are a number of benefits: in fluorescent lamps there is an improvement in efficiency of UV production because of reduction in electrode loss and an increase in PC efficiency; electronic circuits can also provide programmed start and run-up sequences that prolong lamp life; there is no perceptible 50 or 100 Hz flicker from lamps run from electronic circuits at high frequency; and there is no re-strike peak. Figure A1.3.14 shows typical waveforms at 50 kHz. In the case of HP discharges, operation at high frequencies can cause acoustic resonances that result in gross movements or distortions of the arc [10]. The electronic option is then to operate the lamp from a commutated dc—a square wave with fast transition times at frequency in the region of 90–500 Hz. For HP discharges, the lack of flicker and the ability to control lamp power (and thus colour) over life are important benefits.

The optical radiation from discharge sources fluctuates by a percent or two. Part of this is caused by small changes in the cathode termination resulting in arc movement. It has recently been found that

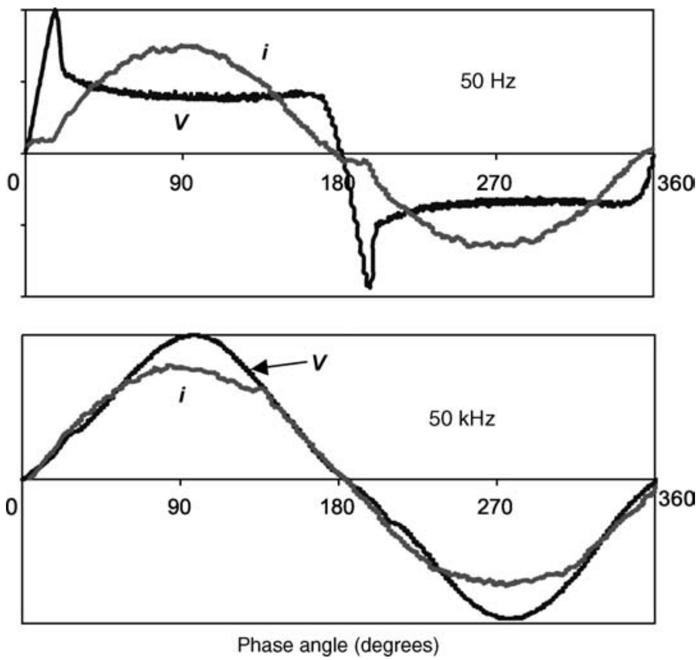


Figure A1.3.14. Measured voltage and current waveforms for a fluorescent lamp operated at frequencies of 50 Hz and 50 kHz. Noise on the waveforms is caused by oscilloscope digitization.

modified square wave supply waveforms can reduce the movement of the arc termination on current reversal [11]. A great improvement in stability can also be achieved by measuring the light output and using it to adjust the power into the lamp (appendix—Oriel Instruments, Light Intensity Control System). A similar device can also be used to control the already excellent stability of tungsten–halogen lamps.

A1.3.9 Other methods of excitation of discharges

A1.3.9.1 Pulsed light sources

A number of lamps are designed for pulsed operation [17]. The obvious example is the xenon flash tube used for photography, laser pumping, as warning beacons and as a transient source for scientific studies. The duration of the flash is of the order of microseconds with repetition frequencies up to hundreds of hertz. Operation is by discharging a capacitor through the lamp. Peak currents may reach thousands of amperes and electrodes must be constructed accordingly. The effects of using pulsed or transient output are that electron temperature can reach substantially higher values than in steady state. The result is usually due to an enhancement of the short wavelength radiation and an increase in peak radiance.

A1.3.9.2 Dielectric barrier discharges

A form of transient discharge, DBDs have been used for large-scale industrial processes such as ozone generation for water purification and for generating far UV radiation for photochemical processes. DBDs may be operated in the pressure range from about 10^2 to 10^5 Pa [16].

Recently a DBD light source, the Osram Planon lamp has been developed. This provides a very uniformly lit tile-shaped area of reasonable luminous efficiency. At present, lamps are made in square format with diagonals up to 540 mm having a uniform luminance of $>6000\,\mathrm{cd\,m^{-2}}$ (appendix—Osram) [14].

The operating principle of a DBD is as follows. High voltage pulses (of some kilovolts) are applied between two electrodes, at least one of which is covered with an insulator of high breakdown strength such as glass. On applying a high voltage pulse, electrons are accelerated towards the anode and form an avalanche that breaches the gap. Electrons arriving at the anode charge up the surface, thus reducing and finally reversing the electric field. Electron current flows first from cathode to anode and then, when the anode charges up, from anode to cathode. The discharge lasts for a time $\sim\mu\mathrm{s}$. During the off period the ionization decays, providing the starting conditions for the next pulse. The discharge therefore comprises a series of micro-discharges with lateral extent approximately equal to the electrode spacing. Microdischarges occur every time the pulse is turned on. DBDs have extremely non-Maxwellian energy distributions in which there are many high energy electrons. Because of this the excitation of resonance states of rare-gas atoms and molecules is favoured, leading to high efficiency of UV production.

The Planon lamp is formed from two glass plates. On the lower plate, a metal cathode interlaced with a metal anode structure is deposited. Both electrodes are coated with glass to form the barrier layers. This form of electrode structure results in very uniform illumination. The lamp is operated from an electronic power supply designed to produce the optimized pulse sequence that is necessary for high efficacy. The two plates are held apart by spacers and the whole structure is sealed and filled with Xe at about 1.4×10^4 (100 Torr). Xe forms an excimer $Xe_2{}^*$ (excited dimer) that radiates efficiently in the vacuum UV at about 172 nm. Phosphor on the inner walls converts the UV to visible radiation. The use of Xe means that the output is almost independent of the lamp temperature so the lamp works just as well outside in cold weather as it does in the confines of office equipment.

The main applications are in displays and office equipment applications where a uniform and high luminance is a requirement. Cylindrical lamps based on the same technology are used in multi-function copiers.

A1.3.9.3 Excitation by induction and by microwaves

In the last decade, a number of inductively coupled lamps have become available commercially from the major lamp manufacturers [9, chapter 11]. All are variants on the fluorescent lamp discharge. Figure A1.3.15 shows a particularly compact example. The coil in the centre is driven at a frequency of about 2.6 MHz. The rate of change of magnetic flux induces a voltage in the azimuthal direction. This causes a current to flow in a torus surrounding the coil. The ballasting is the result of the internal impedance of the supply. Benefits are long life and compactness. Other versions are Philips QL which, with a life of 100 000 h, is designed for use in inaccessible fixtures. Typically these will be high-bay fixtures with lumen packages between 2800 and 9600 lm. The Osram Endura or Icetron lamp which has a stretched torus configuration has higher efficacy and packages of 8000–12000 lm and a rated life of 80000 h.

Microwaves can also be used to excite discharges. Fusion Lighting has pioneered a HP sulphur discharge in which the radiation is emitted by S_2 molecules. The light is white with a CCT in the region of 6000 K and the efficiency of generation can be up to 170 lm/microwave watt—higher than any other white light source. The overall efficiency is reduced because of the relatively poor efficiency of generation of microwave power. Light output levels are very high so the source is used in lighting large buildings. The very high radiance of such sources means that optical means can be used to distribute the light efficiently around buildings.

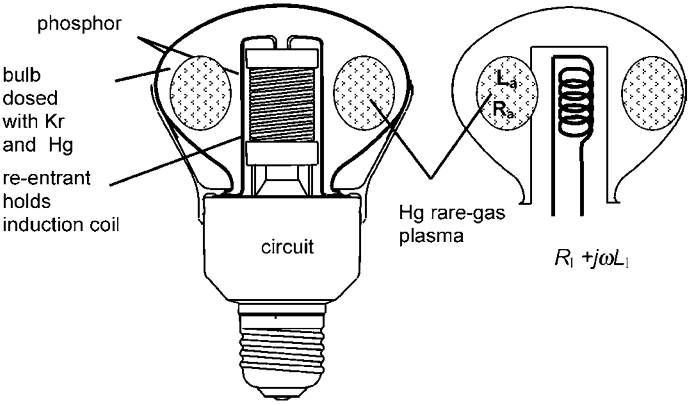


Figure A1.3.15. Inductively coupled discharge lamp (GE Genura). The schematic diagram on the left shows that the plasma is a toroid with inductance L_a and resistance R_a that acts as a secondary to the excitation coil. The primary of the circuit is an impedance $R_l + j\omega L_l$ that includes the effects of the plasma impedance, which depends on the power dissipated in the plasma.

A1.3.10 LEDs from the perspective of conventional lighting

Chapter B1.1 gives detailed information about LEDs. The generation of light by conventional lamps is limited by the black body radiance at the electron temperature. The reason is that the electron motion is randomized. In an LED, the motion of the carriers into the recombination region is far from random; the maximum radiance is therefore not limited by the Planck distribution. Already trichromatic LED [4] assemblies reaching 100 lm W^{-1} have been made. (It is not clear in [24], and in many other LED publications, if this is the flux per lamp watt or flux per wall plug watt.) Note that these lamps are colorimetrically similar to the triphosphor lamps mentioned in section A1.3.3.5. A further advantage of LEDs for many applications is their low étendue (section A1.3.3.3) which means that the light can be directed more efficiently to where it is needed.

LEDs have already made a substantial impact on the conventional lamp manufacturing businesses. There are a number of applications where LEDs are far superior to conventional lamps. The obvious example is traffic signals. The LEDs generate the coloured light only at the wavelengths needed, as compared with the filtered tungsten lamps used until recently. The required signal luminance values can therefore be met at much lower power consumption. It has been estimated that if all traffic lights in the USA were converted to LEDs, electricity consumption would be reduced by 0.4 GW [24]. Moreover, the LEDs are particularly well suited to withstand the vibration they experience as a result of heavy traffic and wind, and they may be frequently switched without damage. But the main advantage is their long life, which dramatically reduces maintenance costs compared with conventional traffic lights. A further advantage is that catastrophic failure of an LED does not have to cause the complete failure of the traffic signal, so safety is improved. All these advantages pay for the extra cost of the LED systems, so we can expect to see a complete takeover of the signal lamp market in land, sea and air transport signs.

White light LEDs are now finding many applications in decorative, aesthetic and artistic lighting where their properties are stimulating designers to produce interesting new ways of using light. There are also other niche markets such as highly localized task lighting where LEDs will make inroads, such as car interiors, desk, stairwell, path lighting etc. In any application requiring relatively low flux levels, conventional lamps can now probably be replaced by LEDs as long as the substitution is not too expensive.

The ultimate target for the LED industry must be the replacement of the huge numbers of fluorescent lamps for lighting offices and other commercial premises. Much is made of the possibility of exceeding the efficiency of present day fluorescent lamps. The future is far from clear however: the issue is the low luminous flux from LEDs. No single LED approaches the level of flux required, so any competitive installation will require many LEDs to produce the required flux levels, although the étendue advantage may reduce the flux levels required for LED installations. But the hard fact is that the cost of making an LED light source increases approximately linearly with flux, whereas the cost of making conventional lamps is very weakly dependent on flux. It is expected that manufacturing costs in both industries will be driven down by competition to comparable levels. If so, replacement of fluorescent lamps by LED equivalents on a global scale may come down to the cost of the materials used to make the lamps. In the case of fluorescent lamps, most of the material cost is in the rare-earth phosphors. Can the semiconductors used in making LEDs ever approach those costs per lumen? Will the potentially low cost organic light emitters ever be able to meet the flux requirements of commercial lighting? We shall see.

References

[1] Bergman R S and Parham T G 1993 Applications of thin film reflecting coating technology to tungsten filament lamps *IEE Proc.* **140A** 418–428

[2] Bloem J, Bouwknegt A and Wesselink G A 1977 Some new mercury alloys for use in fluorescent lamps *J. Illum. Eng. Soc.* 141–147

[3] Boyd R W 1983 *Radiometry and the Detection of Optical Radiation* (New York: Wiley)

[4] Cayless M A 1980 Future developments in lamps *IEE Proc.* **127A** 211–218

[5] Cayless M A and Marsden A M 1983 *Lamps and Lighting* 3rd edn (London: Arnold)

[6] Commision Internationale de l'Éclairage (CIE) 1995 Method of measuring and specifying colour rendering properties of light sources *CIE* 13.3

[7] Commision Internationale de l'Éclairage (CIE) 2001 Colours of light signals *CIE* S004/E:2001

[8] Coaton J R and Fitzpatrick J R 1980 Tungsten–halogen lamps and regenerative mechanisms *IEE Proc.* **127A** 142–148

[9] Coaton J R and Marsden A M 1997 *Lamps and Lighting* 4th edn (the 3rd edition edited by Cayless and Marsden includes more and wider range spectra of lamps) (London: Arnold)

[10] de Groot J J and van Vliet J A J M 1986 *The High-pressure Sodium Lamp* (Deventer: Kluwer)

[11] Derra G, Fischer H E and Mönch 1997 'High pressure lamp operating circuit with suppression of flicker' *US Patent Specification* 5,608,295, 4 March 1997

[12] Elenbaas W 1972 *Light Sources* (London: Macmillan)

[13] Grum F and Becherer R J 1979 *Optical Radiation Measurements: vol 1 Radiometry* (New York: Academic Press)

[14] Hitzschke L and Vollkommer F 2001 *Product Families Based on Dielectric Barrier Discharge: Proc. 9th Int. Symp. on Sci. Tech. Light Sources*, ed R S Bergman (Ithaca, NY: Cornell University Press)

[15] Kelly K L 1963 Line of constant correlated color temperature based on MacAdam's (u,v) uniform transformation of the CIE diagram *J. Opt. Soc. Am.* **53** 999–1002

[16] Kogelschatz U, Eliasson B and Egli W 1999 From ozone generators to flat television screens and future potential of dielectric barrier layer discharges *Pure Appl. Chem.* **71** 1819–1828

[17] Rehmet M 1980 Xenon lamps *IEE Proc.* **127A** 142–148

[18] Richter J 1968 Radiation from hot gases *Plasma Diagnostics*, ed W Lochte-Holtgreven (Amsterdam: North-Holland)

[19] Rutgers G A W and de Vos J C 1954 Relationship between brightness temperature, true temperature and colour temperature of tungsten *Physica* **20** 715–720

[20] Stupp E H and Brennesholtz 1999 *Projection Displays* (New York: John Wiley)

[21] Waymouth J F 1971 *Electric Discharge Lamps* (Cambridge: MIT Press)

[22] Weston G F 1968 *Cold Cathode Discharge Tubes* (London: Iliffe)

[23] Wyszecki G and Stiles W S 2000 *Color Science: Concepts and Methods, Quantitative Data and Formulae* (New York: Wiley)

[24] Zukauskas A, Shur M S and Caska R 2002 *Introduction to Solid-state Lighting* (New York: Wiley)

Appendix. Selected manufacturers and suppliers of lamps

The manufacturers on this list give particularly helpful data in catalogues and/or websites for lamps with actual or potential electro-optical applications.

Cathodeon	lamps for scientific instruments
Ealing Electro-Optics	lamp units for integration into optical systems
Fusion Lighting	microwave discharge lamps
GE Lighting	full range of lamps for illumination and special purposes
Harrison Electrical	cold cathode fluorescent
Heraeus Noblelight	special lamps mainly for industrial and scientific processes
Iwasaki	full range of lamps for illumination and special purposes
Osram	full range of lamps for illumination and special purposes
Oriel Instruments	lamp units for integration into optical systems, spectra of lamps
Philips Lighting	full range of lamps for illumination and special purposes
Stanley	cold cathode fluorescent
Toshiba Lighting	full range of lamps for illumination and special purposes
Ushio	wide range lamps for audio-visual, entertainment, photographic, scientific/media and industrial processes
Welch Allyn	lamps for special applications

A1.4

Detection of optical radiation

Antoni Rogalski and Zbigniew Bielecki

The birth of photodetectors can be dated back to 1873 when Smith discovered photoconductivity in selenium. Progress was slow until 1905, when Einstein explained the newly observed photoelectric effect in metals, and Planck solved the blackbody emission puzzle by introducing the quanta hypothesis. Applications and new devices soon flourished, pushed by the dawning technology of vacuum tube sensors developed in the 1920s and 1930s culminating in the advent of television. Zworykin and Morton, the celebrated fathers of videonics, on the last page of their legendary book *Television* (1939) concluded that: "when rockets will fly to the moon and to other celestial bodies, the first images we will see of them will be those taken by camera tubes, which will open to mankind new horizons". Their foresight became a reality with the Apollo and Explorer missions. Photolithography enabled the fabrication of silicon monolithic imaging focal planes for the visible spectrum beginning in the early 1960s. Some of these early developments were intended for a picturephone, other efforts were for television cameras, satellite surveillance, and digital imaging. Infrared imaging has been vigorously pursed in parallel with visible imaging because of its utility in military applications. More recently (1997), the CCD camera aboard the Hubble space telescope delivered a deep-space picture, a result of 10-days integration, featuring galaxies of the 30th magnitude—an unimaginable figure even for astronomers of our generation. Probably, the next effort will be in the big-band age. Thus, photodetectors continue to open to mankind the most amazing new horizons.

Before proceeding to the detailed description of detection of optical radiation, it is now appropriate to digress on system considerations concerning photodetection. We would like to determine how good performance is in view of fundamental limits of sensitivity and speed of response, irrespective of the actual type of detector used. Next, different application circuits used in direct detection systems together with elucidation of the design of front-end circuits and discussion of their performance are presented. The third part of the chapter is devoted to advanced techniques in photodetection covering topics not usually found in textbooks and demonstrating how photodetection is far from being a completely explored field. In the last part, the updated information devoted to readout of signals from detector arrays and focal plane arrays (FPAs) is included. It is shown that detector focal plane technology has revolutionized many kinds of imaging in the past 25 years.

A1.4.1 Detection regimes and figures of merit

A common problem of any type of photon detector (yielding emitted electrons or internal electron–hole pairs as a response to incoming photons) is how to terminate the photodetector with a suitable load resistor, and to trade off the performance between bandwidth and signal-to-noise ratio. This is necessary for a wide family of detectors, including phototubes, photoconductors, photodiodes, CCDs, vidicon targets, etc, all of which are described by a current generator I_{ph} with a stray capacitance C across it.

Let us consider the equivalent circuit of a photodetector ending on a load resistor R_{L}, as shown in figure A1.4.1. We indicate current noise generator with diamond shape and asterisk. This is the basic circuit for detection, but, as will be shown later, it is not the best one to give a good compromise between bandwidth and noise. The output signal in voltage $V = IR_{\mathrm{L}}$ or in current I, has a bandwidth, or 3 dB high-frequency cutoff given by

$$\Delta f = \frac{1}{2\pi R_{\mathrm{L}} C}. \tag{A1.4.1}$$

Two noise contributions are added to the signal. One is the Johnson (or thermal) noise of the resistance R_{L}, with a quadratic mean value

$$I_{\mathrm{nR}}^2 = \frac{4kT\Delta f}{R_{\mathrm{L}}} \tag{A1.4.2}$$

where k is the Boltzmann constant and T is the absolute temperature.

The total current $I = I_{\mathrm{ph}} + I_{\mathrm{d}}$ is the sum of the signal current and the dark current. With this current is associated the quantum (or shot) noise arising from the discrete nature of electrons and photoelectrons. Its quadratic mean value is given by

$$I_{\mathrm{n}}^2 = 2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f \tag{A1.4.3}$$

where q is the electron charge and Δf is the observation bandwidth, as in equation (A1.4.1).

A general noise equivalent circuit for a photodetector is shown in figure A1.4.1(*b*). The above two fluctuations are added to the useful signal and the corresponding noise generators are placed across the device terminals. Since the two noises are statistically independent, it is necessary to combine their quadratic mean values to give the total fluctuation as

$$I_{\mathrm{n}}^2 = 2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f + \frac{4kT\Delta f}{R_{\mathrm{L}}}. \tag{A1.4.4}$$

From equations (A1.4.1) and (A1.4.4) it can be seen that bandwidth and noise optimization impose opposite requirements on the value of R_{L}. To maximize Δf one should use the smallest possible R_{L}, whilst to minimize I_{n}^2 the largest possible R_{L} is required. A photodetector can have a good sensitivity using very high load resistances (up to GΩs), but then only modest bandwidths (≈ kHz or less), or it can be made fast by using low load resistances (e.g. $R_{\mathrm{L}} = 50\,\Omega$) but at the expense of sensitivity.

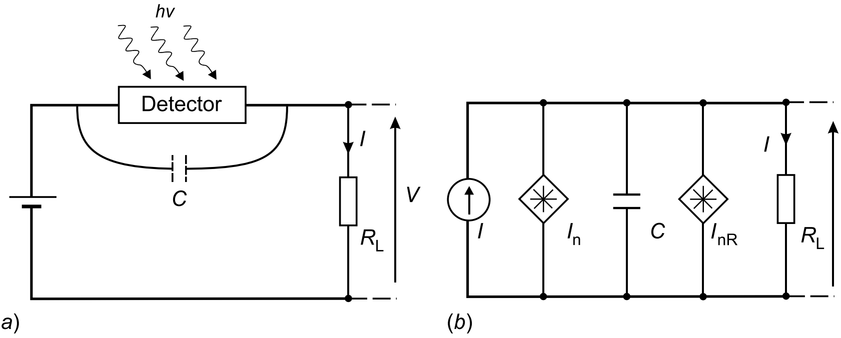


Figure A1.4.1. General circuit (*a*) of a photodetector and its equivalent circuit (*b*) including noise generators (noise generators shown with asterisk).

Using equation (A1.4.4), we can evaluate the relative weight of the two terms in total noise current. In general, the best possible sensitivity performance is achieved when the shot noise is dominant compared to the Johnson noise, i.e.

$$2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f \geq \frac{4kT\Delta f}{R_{\mathrm{L}}}$$

which implies the condition

$$R_{\mathrm{Lmin}} \geq \frac{2kT/q}{I_{\mathrm{ph}} + I_{\mathrm{d}}}. \tag{A1.4.5}$$

From this equation at room temperature, then

$$R_{\mathrm{L\,min}} \geq \frac{50\,\mathrm{mV}}{I_{\mathrm{ph}} + I_{\mathrm{d}}}.$$

So, at low signal levels (for $I_{\mathrm{ph}} \ll I_{\mathrm{d}}$), very high values of resistance are required; for example, for $I_{\mathrm{d}} = 5\,\mathrm{pA}$, not an unusually low dark current, then $R_{\mathrm{Lmin}} = 10\,\mathrm{G\Omega}$. However, if we are using a resistor termination value $R_{\mathrm{L}} < R_{\mathrm{Lmin}}$ then, the total noise can be written as

$$I_{\mathrm{n}}^2 = 2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f\left(1 + \frac{R_{\mathrm{Lmin}}}{R_{\mathrm{L}}}\right). \tag{A1.4.6}$$

This means that the noise performance is degraded by factor $R_{\mathrm{Lmin}}/R_{\mathrm{L}}$, compared to the intrinsic limit allowed by the dark current level. Thus, using $R_{\mathrm{L}} < R_{\mathrm{Lmin}}$ means that the shot-noise performance is reached at a level of current not less than

$$I_{\mathrm{ph}} + I_{\mathrm{d}} \geq \frac{2kT/q}{R_{\mathrm{L}}}. \tag{A1.4.7}$$

From this equation, we see that for a fast photodiode with a $50\,\Omega$ load at $T = 300\,\mathrm{K}$, this current has very large value of 1 mA.

The above considerations are valid for photodetectors without any internal gain, G. Let us now extend the calculation of the signal-to-noise ratio (S/N) to photodetectors having internal gain. In this case, the shot noise can be expressed in the form

$$I_{\mathrm{n}}^2 = 2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f G^2 F \tag{A1.4.8}$$

where F is the excess noise factor to account for the extra noise introduced by the amplification process. Of course, in a non-amplified detector, $F = 1$ and $G = 1$. The total noise is a sum of shot noise and thermal noise

$$N = \left[2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f G^2 F + \frac{4kT\Delta f}{R_{\mathrm{L}}}\right]^{1/2} \tag{A1.4.9}$$

and then we obtain a S/N ratio

$$\frac{S}{N} = \frac{I_{\mathrm{ph}}}{\left[2q(I_{\mathrm{ph}} + I_{\mathrm{d}})\Delta f F + (4kT\Delta f/R_{\mathrm{L}}G^2)\right]^{1/2}}. \tag{A1.4.10}$$

If we now introduce a critical value I_{ph0}, called the *threshold of quantum regime*

$$I_{\mathrm{ph0}} = I_{\mathrm{d}} + \frac{2kT/q}{R_{\mathrm{L}}FG^2}$$

then, equation (A1.4.10) becomes

$$\frac{S}{N} = \frac{I_{\mathrm{ph}}}{[2q(I_{\mathrm{ph}} + I_{\mathrm{ph0}})\Delta fF]^{1/2}}. \tag{A1.4.11}$$

Analysing equation (A1.4.11), two detection regimes can be found, according to whether the signal I_{ph} is larger or smaller than I_{ph0}. For the signals, $I_{\mathrm{ph}} > I_{\mathrm{ph0}}$, and $F = 1$

$$\frac{S}{N} = \left(\frac{I_{\mathrm{ph}}}{2q\Delta f}\right)^{1/2}. \tag{A1.4.12}$$

This *S/N* is called the *quantum noise limit* of detection. This limitation cannot be overcome by any detection system, whether operating on coherent or incoherent radiation. In fact, equation (A1.4.12) is a direct consequence of the quantitative nature of light and the Poisson photon arrival statistics.

In the small signal regime, $I_{\mathrm{ph}} < I_{\mathrm{ph0}}$, we have

$$\frac{S}{N} = \frac{I_{\mathrm{ph}}}{(2qI_{\mathrm{ph0}}\Delta f)^{1/2}}. \tag{A1.4.13}$$

That is, the S/N ratio is proportional to the signal, and the noise has a constant value, primarily given by the load resistance. This is the *thermal regime of detection*.

Figure A1.4.2 shows the trend of the S/N ratio (standardized to $(2q\Delta f)^{1/2}$), as a function of the signal amplitude $I_{\mathrm{ph}}/I_{\mathrm{ph0}}$. We can notice that in the thermal regime, the slope is 20 dB per decade up to the threshold $I_{\mathrm{ph}}/I_{\mathrm{ph0}} = 1$, and from here onward the slope becomes 10 dB per decade in the quantum regime. The effect of an excess noise factor F is also shown in figure A1.4.2.

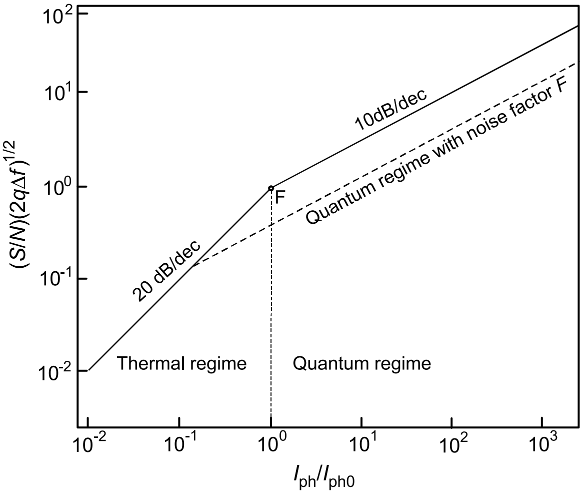


Figure A1.4.2. The S/N ratio of a photodetector, as a function of the input signal, in the thermal and quantum regimes of detection. Reproduced from [1].

The threshold of the quantum regime $I_{\mathrm{ph}} = I_{\mathrm{ph0}} = I_{\mathrm{d}} + (2kT/q)/R_{\mathrm{L}}FG^2$ is the signal level that corresponds to the break point between thermal and quantum regimes. When the dark current is very small, the second term is the dominant one; any eventual internal gain greatly helps because it scales the load resistance as G^2.

In all cases, I_{ph0} can be interpreted as the equivalent dark current level of the photodetector and has value of

$$I_{\mathrm{ph0}} = \frac{2kT/q}{R_{\mathrm{eq}}G^2} \tag{A1.4.14}$$

with

$$\frac{1}{R_{\mathrm{eq}}} = \frac{1}{R_{\mathrm{F}}} + \frac{q}{2kT}G^2 I_{\mathrm{d}}.$$

We define R_{eq} as the noise-equivalent load resistance of the photodetector.

To provide ease of comparison between detectors, certain figures of merit, computed from the measured data, have been defined.

The voltage (or analogous current) responsivity is given by

$$R = \frac{Q_{\mathrm{u}}}{P} \tag{A1.4.15}$$

where Q_{u} is the output quantity supplied by the detector (e.g. a current I_{u}, a voltage V_{u}, or any other physical quantity) and P is the incident radiant power.

At equal responsivity, the detector with the smallest output noise Q_{u} on the useful signal is the most sensitive. Therefore, the first figure of merit for a detector is the NEP—noise equivalent power—defined as the ratio of output noise to responsivity:

$$\mathrm{NEP} = \frac{g_{\mathrm{n}}}{R}. \tag{A1.4.16}$$

So, the NEP represents the input power that gives a unity signal to noise ratio, $S/N = 1$ at the output; that is, a marginal condition of detection.

The better the detector performance is, since the smaller the NEP is. Therefore, it is more convenient to define its inverse as a merit figure. In addition, it should be taken into consideration that whatever the noise source is, it can be expected that the noise quadratic total value is proportional to observation bandwidth Δf and detector area A. Thus it is even better to take, as the intrinsic noise parameter of a detector, the ratio $\mathrm{NEP}/(A\Delta f)^{1/2}$ normalized to unit area and bandwidth. In order to simplify the comparison of different detectors and to have a parameter that increases as the performance improves, the detectivity D^* (called *D-star*) is defined as:

$$D^* = \frac{(A\Delta f)^{1/2}}{\mathrm{NEP}}. \tag{A1.4.17}$$

This is the fundamental figure of merit used for detectors. It can be transformed to the following equation

$$D^* = \frac{(A\Delta f)^{1/2}}{P}\frac{S}{N}. \tag{A1.4.18}$$

D^* is defined as the rms signal-to-noise ratio (S/N) in a 1 Hz bandwidth, per unit rms incident radiation power, per square root of detector area. D^* is expressed in cm $\mathrm{Hz}^{1/2}\,\mathrm{W}^{-1}$, a unit which has recently been called a ‘Jones’.

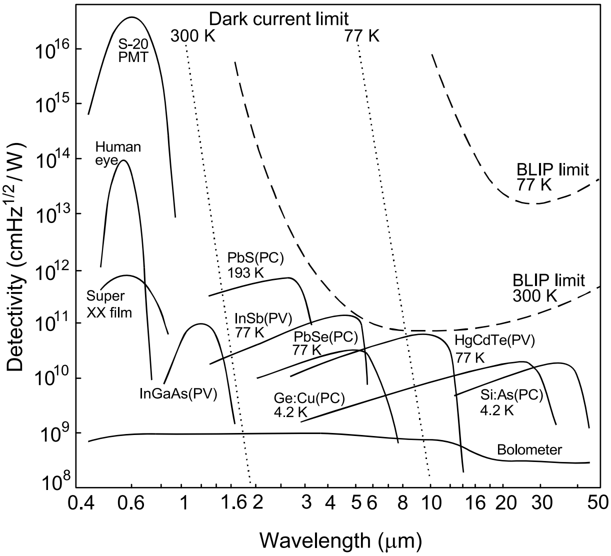


Figure A1.4.3. Detectivity as a function of wavelength for a number of different photodetectors. The BLIP and the dark current limits are indicated. PC—photoconductive detector and PV—photovoltaic detector, PMT—photomultiplier tube.

As mentioned already, the ultimate performance of infrared detectors is reached when the detector and amplifier noise are low compared to the photon noise. The photon noise is fundamental, in the sense that it arises not from any imperfection in the detector or its associated electronics but rather from the detection process itself, as a result of the discrete nature of the radiation field. The radiation falling on the detector is a combination of that from the target (signal) and that from the background. The practical operating limit for most infrared detectors is the background fluctuation limit, also known as the background limited infrared photodetector (BLIP) limit.

Figure A1.4.3 compares typical D^*s of different detectors as a function of wavelength. Also the BLIP and the dark current limits are indicated.

A1.4.2 Direct detection systems

A1.4.2.1 Introduction

A receiver of optical radiation consists of a photodetector, preamplifier, and signal processing circuit (figure A1.4.4). In a photodetector, the optical signal is converted to an electrical one, which is amplified before further processing. The sensitivity of an optical detection systems depends primarily on the first stage of a photoreceiver, i.e. the photodetector and preamplifier.

A preamplifier should have low noise and a sufficiently wide bandwidth to ensure faithful reproduction of the temporal shape of an input signal. Figure A1.4.4 shows the so-called, direct detection system. It is necessary to minimize the noises from various sources, i.e. background noise, photodetector noise, biasing resistors noise, and any additional noises of signal processing. If further minimization of noise of the first photoreceiver stages is not possible, advanced methods of optical signal detection can sometimes be used to still recover information carried by optical radiation signals of extremely low power. Heterodyne and homodyne detection can be used to reduce the effects of amplifier noise. Post detection methods are: phase-sensitive detection and synchronous integration of a signal.

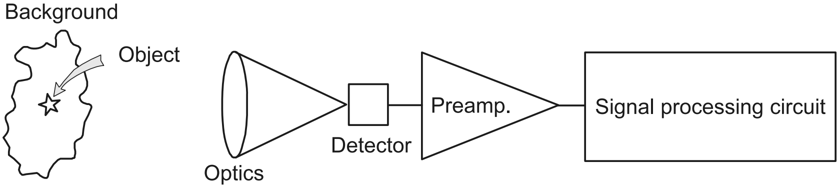


Figure A1.4.4. Block diagram of an optical radiation receiver.

A1.4.2.2 Selection of active amplifying elements

Several types of discrete devices or ICs are suitable for the active element in preamplifiers: bipolar (BJT) or field-effect transistor (FET) or an integrated circuit (IC) with an input bipolar, FET or MOSFET transistor can be used.

The most important parameter of each receiving device is its signal-to-noise ratio (S/N). Because low-level signals reach the photoreceiver, the noise optimization of a system, i.e. to obtain maximum S/N is a very important problem [2]. Optimum design of a preamplifier can be obtained by analysis of particular noise sources in a detector–preamplifier circuit. The equivalent input noise V_{ni} will be used to represent all noise sources. A scheme of detector–preamplifier noise circuit is shown in figure A.1.4.5. A level of equivalent noise at the input of this circuit is determined by the detector noise V_{nd}, the background noise V_{nb}, and the preamplifier noise. The preamplifier noise is represented completely by the zero impedance voltage generator V_{n} in series with the input port, and the infinite impedance current generator I_{n} in parallel with the input. Typically, each of these terms is frequency dependent. For non-correlated noise components, the equivalent noise at the input of a photodetector–preamplifier circuit is described by the formula

$$V_{\mathrm{ni}}^2 = V_{\mathrm{nd}}^2 + V_{\mathrm{nb}}^2 + V_{\mathrm{n}}^2 + I_{\mathrm{n}}^2 R_{\mathrm{d}}^2 \tag{A1.4.19}$$

where R_{d} is the detector resistance.

This single noise source, located at V_{s}, can be substituted for all sources of the system noise. Note that V_{ni}^2 is independent of the amplifier gain and its input impedance. Thus, V_{ni}^2 is the most useful index against which the noise characteristics of various amplifiers and devices can be compared.

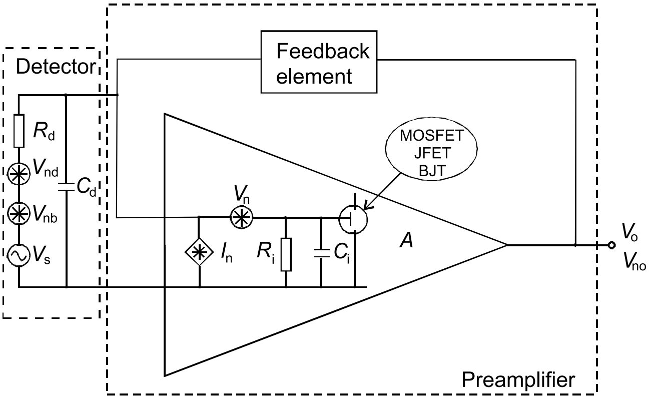


Figure A1.4.5. Noise equivalent diagram of a photodetector–preamplifier circuit; R_{i} and C_{i} is the input resistance and the capacity of a preamplifier, respectively, A is the voltage gain of the preamplifier.

The first transistor of a preamplifier is a dominant noise contributor of the input noise of signal processing circuits (figure A1.4.5). The effective contribution of this transistor noise also depends on the detector impedance. If the transistor of the input stage has a high noise current, it will be a bad choice for use with a high-resistance (current source) detector. A low-resistance detector operates best with a preamplifier of low noise voltage.

Figure A1.4.6 shows the dependence of the ratio of preamplifier (or input transistor) noise to detector thermal noise as a function of detector resistance, for bipolar, JFET, and MOSFET amplifiers, in a common emitter or common source configuration [3]. It can be noticed that there are some ranges of detector resistance for which the preamplifier noise is lower than the detector thermal noise. For any given detector, the values of the thermal noise voltage V_t and the resistance R_d are determined by the detector type (they cannot be changed). However, it is possible to change the parameters V_n and I_n of the designed preamplifier. Thus, minimization of the input noise of the circuit is possible. Changes in the values V_n and I_n are made by choosing adequate elements in the preamplifier circuit. For low-resistance detectors (from tens Ω to 1 kΩ), circuits with bipolar transistors at the input are usually used as they have low values of V_n. Sometimes, in order to decrease the value of optimal resistance R_o, a parallel connection of several active elements or even parallel connection whole amplifiers is advantageous [4]. Within the range of average detector resistances from 1 kΩ to 1 MΩ, the preamplifiers FET input stages can be used. However, for connection with high-resistance detectors (above 1 MΩ) especially recommended are the transistors of low I_n values. Such requirements fulfil JFET and MOSFET transistors.

A1.4.2.3 First stages of photoreceivers

There are two general types of photon detectors without internal gain: photoconductive and junction devices (photovoltaic ones). These photodetectors are used with many types of preamplifiers. The choice of circuit configuration for the preamplifier is largely dependent upon the system application. The two basic preamplifier structures—voltage amplifiers type and transimpedance (current in-voltage out type) will be discussed. The voltage preamplifiers can be either low impedance or high impedance ones. A simplified circuit diagram is shown in figure A1.4.7. Bias voltage is supplied by V_b. The detector signal is

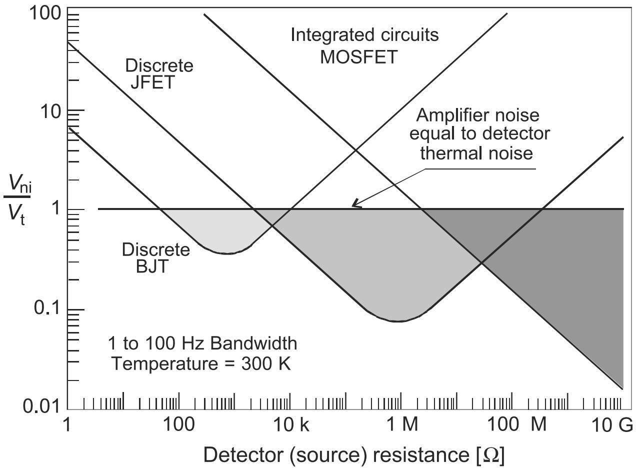


Figure A1.4.6. Dependence of the ratio of preamplifier input voltage noise to detector thermal noise, as a function of detector resistance. Reproduced from [3].

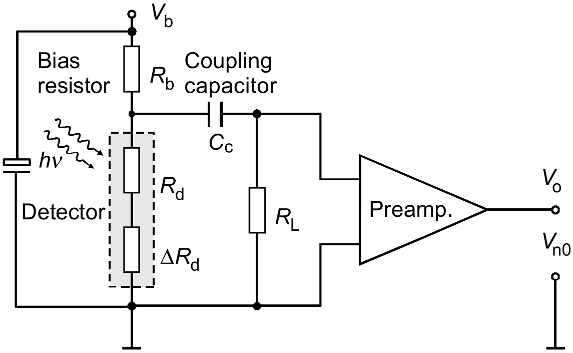


Figure A1.4.7. A simple circuit of voltage mode preamplifier.

developed as a voltage drop across R_b. The variable resistance component of the detector R_d is represented by the incremental resistance ΔR_d. The load resistance R_L provides a bias path for the amplifier input.

The simplest preamplifier structure is the low input impedance voltage preamplifier. This design is usually implemented using a bipolar transistor. Either common emitter or grounded emitter input stages may be designed with a reasonably low input impedance. In the low input impedance preamplifiers, the signal source is loaded with a low impedance (e.g. 50 Ω) input stage. The time constant of the temporal response is determined by the combined load resistance and input capacitance of the detector and preamplifier and this determines the detection bandwidth. Preamplifiers of low input resistance can provide high bandwidth but not very sensitive photoreceiver. Photoreceivers with preamplifiers of low input resistance are, therefore, usually used when wide transmission band is required.

The high-impedance preamplifier gives a significant improvement in sensitivity over the low-impedance preamplifier, but it requires considerable electronic frequency equalization to compensate for its high-frequency roll-off. The preamplifier has problems of limited dynamic range, being easily saturated at higher input power levels. When a highly sensitive photoreceiver with a low dynamic range is needed, preamplifier of high input impedance is recommended. However, for this, there is a better configuration, that we shall now discuss. This is called the transimpedance receiver.

The transimpedance preamplifier finds many applications in optical signals detection. A schematic of this is shown in figure A1.4.8. In this circuit, R_d is the detector resistance, and it can be a photodiode or a photoresistor. Depending on detector type and required application, the detector can be biased from V_b or connected directly across the input without bias. The current produced by the detector flows through the resistor R_f located in a feedback loop. The optional potentiometer R_b is used for setting the zero value of the output voltage without detector illumination [5]. It can be used for compensation of a dc level at the output of the circuit resulting from background radiation. Usually, the input bias current is negligible, so $I_{ph} = -I_f$.

The voltage at the preamplifier output is given by the formula

$$V_o = -\frac{R_f}{R_d} V_b = -I_{ph} R_f \tag{A1.4.20}$$

proper for low frequencies. The great advantage of this configuration is that the preamplifier gives low noise performance without the severe limitation on bandwidth imposed by high input impedance preamplifier. The bandwidth is much higher because the effective load resistance is very low, and has value $R_f/(A+1)$ where A is the gain of the amplifier at the appropriate frequency. It also provides greater

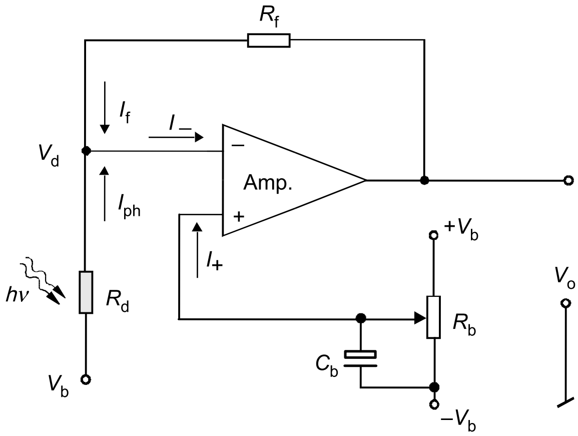


Figure A1.4.8. Schematic diagram of transimpedance amplifier.

dynamic range than the high input impedance structure. It is possible to linearly detect and process optical signals many orders of powers magnitude.

At low frequencies, using a photodiode in unbiased configuration, across the input of transimpedance preamplifier, avoids the noise and dc offset problems that dark current might otherwise cause (non-inverting input is on the ground). In this configuration, there is the small price of increased photodiode capacitance, which would be far more serious at high frequencies. For InGaAs photodiodes can have substantially lower quantum efficiencies if unbiased (due to much thinner depletion region) although the benefit to avoid dark current can be more useful here.

The effective input impedance R_i at the amplifier input as a result of the effects of the feedback circuit is given simply by

$$R_i = \frac{R_f}{A(f)+1} \approx \frac{R_f}{A(f)} \tag{A1.4.21}$$

where $A(f)$ is the frequency-dependent gain of the preamplifier. As this input impedance is effectively in parallel with the capacitance $(C_d + C_i)$, the 3 dB breakpoint $f_{-3\,\mathrm{dB}}$ in the frequency response is given by

$$f_{-3\,\mathrm{dB}} = \frac{1}{2\pi R_i(C_d + C_i)} = \frac{A(f)}{2\pi R_f(C_d + C_i)}. \tag{A1.4.22}$$

If $A(f)$ is described in terms of gain–bandwidth product $A(f)\Delta\mathrm{f}$ (or GBP) such that GBP $= \Delta f A(f)$, then

$$f_{-3\,\mathrm{dB}} = \frac{\mathrm{GBP}/f_{-3\,\mathrm{dB}}}{2\pi R_f(C_d + C_i)} \tag{A1.4.23}$$

therefore

$$f_{-3\,\mathrm{dB}} = [\mathrm{GBP}/2\pi R_f(C_d + C_i)]^{1/2}. \tag{A1.4.24}$$

A practical problem with many transimpedance designs is to achieve a sufficiency high open-loop gain product, to keep the $f_{-3\,\mathrm{dB}}$ value high, yet still maintain a sufficiently low open-loop phase shift to maintain stability. It often proves necessary, in practice, to use a value of R_L other than desired on low-noise grounds in order to achieve the necessary $f_{-3\,\mathrm{dB}}$. This gives rise to the widely held belief that transimpedance designs are inherently noisier than high impedance input design. However, this belief is not valid with good high-gain bandwidth designs. Secondly, even with sub-optimum amplifier design, achieving poorer gain bandwidth, it is always possible to perform an engineering compromise, by choosing a high value of R_L, to maintain low-noise, and use small amount of equalization after the transimpedance amplifier to compensate for the resulting roll off.

The preamplifiers described earlier are used with semiconductor detectors. Now, we would like to present exemplary preamplifiers to the photomultiplier tubes.

The operating principles, construction and characteristics of a photomultiplier are given in section B2. Now, we would like to analyse matching the preamplifier to the photomultiplier. Different pulse processing techniques are typically employed, depending on whether the arrival time or the amplitude (energy) of the detected event must be measured. Three basic types of preamplifiers are available: the current-sensitive preamplifier, the parasitic-capacitance preamplifier, and the charge-sensitive preamplifier. The simplified schematic of the current-sensitive preamplifier is shown in figure A1.4.9. The $R = 50\,\Omega$ input impedance of the current-sensitive preamplifier provides proper termination of the $50\,\Omega$ coaxial cable, and converts the current pulse from the detector to a voltage pulse. The amplitude of the voltage pulse at the preamplifier output will be

$$V_o = RI_iA \tag{A1.4.25}$$

where I_i is the amplitude of the current pulse from the detector.

For counting applications this signal can be fed to a fast discriminator, the output of which is recorded by a counter-timer. For timing application, the dominant limitation on timing resolution with photomultiplier tubes is fluctuation in the transit times of the electrons. This causes a jitter in the arrival time of the pulse at the detector output. If the detector signals are small enough to require a current-sensitive preamplifier, the effect of a preamplifier input noise on time resolution must also be considered. It is important to choose a current-sensitive preamplifier whose rise time is much faster than detector rise time. Unfortunately, the faster preamplifier does contribute extra noise, because of the unnecessarily wide bandwidth. The excess noise will increase the timing jitter. Choosing a preamplifier rise time that is much slower than the detector rise time reduces the preamplifier noise contribution but causes degradation in pulse rise time and its amplitude. Consequently, the timing jitter becomes worse. The optimum choice depends on the rise time and amplitude of the detector signal. Most current-sensitive preamplifiers designed for timing applications have ac-coupled.

The most effective preamplifier for these detectors is the parasitic-capacitance preamplifier shown in figure A1.4.10. It has a high input impedance (above $1\,M\Omega$).

The parasitic capacitance is presented by the detector and the preamplifier input (typically 10–50 pF). The resulting signal is a voltage pulse having an amplitude proportional to the total charge in the detector pulse. This type of preamplifier is sensitive to small changes in the parasitic capacitance. The rise time is equal to the duration of the detector current pulse. A resistor connected in parallel with the input capacitance causes an exponential decay of the pulse. An amplifier-follower is included

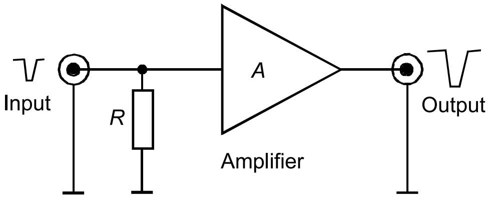


Figure A1.4.9. A simplified schematic of the current-sensitive preamplifier.

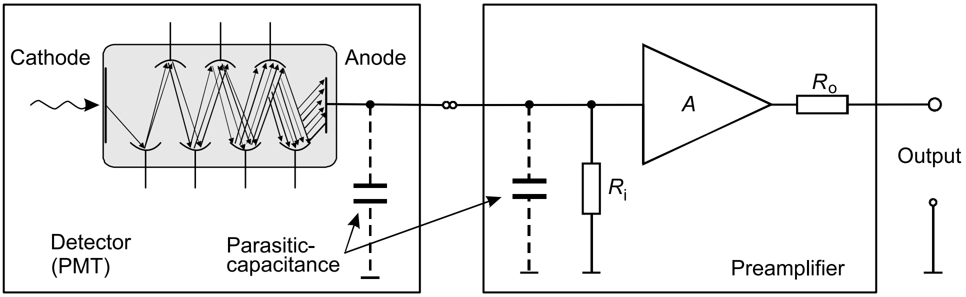


Figure A1.4.10. A simplified diagram of the parasitic-capacitance preamplifier.

as a buffer to drive the low impedance of a coaxial cable at the output. These preamplifiers are highly recommended for photomultiplier tubes, microchannel plate MPTs, and scintillation detectors.

For most energy spectroscopy applications a charge-sensitive preamplifier is preferred (figure A1.4.11). This preamplifier integrates the charge on the feedback capacitor. Its gain is not sensitive to a change in detector capacitance. The output voltage from the preamplifier (V_o) and the decay time constant (τ_f) are given respectively by

$$V_o = \frac{Q_d}{Q_f} = \frac{Eq}{\varepsilon C_f} 10^6 [\mathrm{mV}] \quad \text{and} \quad \tau_f = R_f C_f \tag{A1.4.26}$$

where Q_d is the charge of the detector, C_f is the feedback capacitor, E is the energy in MeV of the incident radiation, q is the charge of an electron, ε is the amount of energy (eV) required to produce an electron–hole pair in the detector, and 10^6 converts MeV to eV. The resistor R_f should be made consistent as much as possible with the signal energy-rate product and the detector leakage current. The input capacitance must be much greater than the other sources of capacitance connected to the preamplifier input in order for the preamplifier sensitivity to be unaffected by external capacitance changes. The stability of the preamplifier sensitivity is dependent on the stability of the feedback capacitor (C_f is selected for good temperature stability) and the preamplifier open loop gain. The open loop gain will be very large so that small changes in the C_f can be neglected.

The rise time of the output pulse of the preamplifier, in the ideal case, is equal to the charge collection time of the detector. When detectors with very fast collection times or large capacitances are used, the preamplifier itself may limit the rise time of V_o.

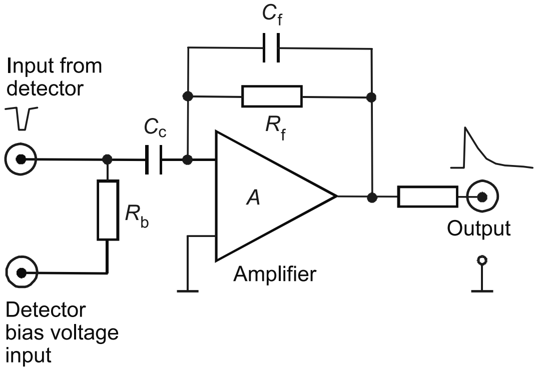


Figure A1.4.11. A simplified diagram of the charge-sensitive preamplifier.

A1.4.2.4 Photon-counting techniques

Photomultiplier tubes are also used in photon counting. Photon-counting is one effective way to use a photomultiplier tube for measuring very low light (e.g. astronomical photometry and fluorescence spectroscopy). A number of photons enter the photomultiplier tube and create an output pulse signal (figure A1.4.12(*a*)). The actual output pulse obtained by the measurement circuit is a dc with fluctuation (figure A1.4.12(*b*)).

When the light intensity becomes so low that the incident photons are separated as shown in figure A1.4.12(*c*), this condition is called a single photon event. The number of output pulses is in direct proportion to the amount of incident light and this pulse counting method is advantageous for signal-to-noise ratio and stability over the dc method averaging all the pulses. This counting technique is called the photon-counting method.

Since the photomultiplier tube output contains a variety of noise pulses in addition to signal pulses representing photoelectrons, simply counting of the pulses, without some form of noise elimination, will not result in an accurate measurement. The most effective approach to noise elimination is to investigate the height of the output pulses. Figure A1.4.13 shows the output pulse and discriminator level.

A typical pulse height distribution (PHD) of the output of the photomultiplier tubes is shown in figure A1.4.14. In this PHD method, the low level discrimination (LLD) is set at the valley and the upper level discrimination (ULD) at the foot. Most pulses smaller than the LLD are noises and pulses larger then ULD result from interference (e.g. cosmic rays). Therefore, by counting the pulses between the LLD and UPD, accurate light measurements are made possible. In the PHD, H_m is the mean height of pulses. It is recommended that the LLD be set at 1/3 of H_m and the ULD a triple H_m.

In addition, the avalanche photodiodes (APDs) can get output pulses for each detected photon and thus potentially very high sensitivities, comparable to that of photomultipliers. They are selected based on having extremely low noise and low bulk dark-current. They are intended for ultra-low light level application (optical power less than 1 pW) and can be used in either their normal linear mode (polarization voltage $V_R <$ breakdown voltage V_{BR}), or as photon counters in the 'Geiger' mode ($V_R > V_{BR}$) where a single photoelectron may trigger an avalanche pulse of about 10^8 carriers.

In the linear mode operation, the APD is well suited for application which requires high sensitivity and fast response time; for example: laser rangefinders, fast receiver modules, lidar, and ultrasensitive spectroscopy.

When biased above the breakdown voltage, an avalanche photodiode will normally conduct a large current. However, if this current is limited to less than the APD's 'latching' current, there is a strong statistical probability that the current will fluctuate to zero in the multiplication region, and the APD will then remain in the 'off' state until an avalanche pulse is triggered by either a bulk or photo-generated carrier. If the number of bulk carrier generated pulses is low, the APD can therefore, be used to count individual current pulses from incident photons. The value of the bulk dark current is, therefore, a

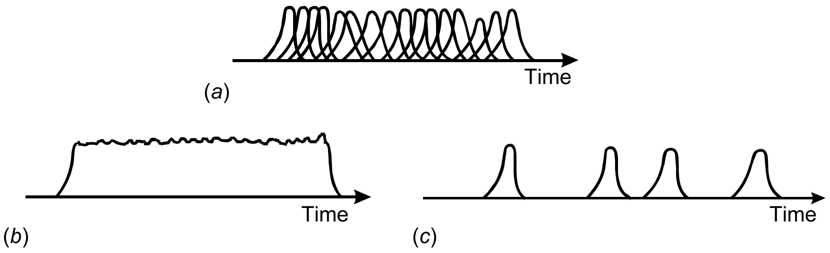


Figure A1.4.12. Pulse signals at the output of a photomultiplier tube.

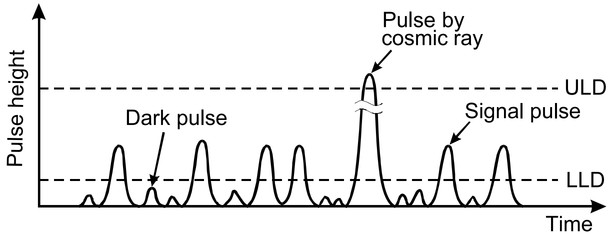


Figure A1.4.13. Output pulse and discriminator level.

significant parameter in selecting an APD for photocounting, and can be reduced exponentially by cooling.

The APDs can be used in the Geiger mode using either 'passive' or 'active' pulse quenching circuits.

A passive quench takes the current from the diode and passes it through a load resistor and a series resistor, causing the bias voltage to drop. An active quench uses a transistor to lower the bias voltage. Passive quench circuits, although simple, limit the rate at which photons can be counted, because time has to be allowed for the quench and for the return of the voltage to break down before the next photon arrives. This characteristic time is basically given by RC, where R is the series resistance and C is device capacitance. Because R must be sufficiently large to trigger the quench, the time is limited by device capacitance to typically hundreds of nanoseconds, and time resolution is at best about 400 ps.

The simplest, and in many cases a perfectly adequate method of quenching a breakdown pulse, is through the use of a current-limiting load resistor. An example of such a passive quenching circuit is shown in figure A1.4.15(*a*). The load-line of the circuit is shown in figure A1.4.15(*b*).

To be in the conducting state at V_{BR}, two conditions must be met:

- the avalanche must have been triggered by either a photoelectron or a bulk-generated electron entering the avalanche region of the APD, to continue to be in the conducting state, and
- a sufficiently large current, called latching current must be passing through the device so there is always an electron or hole in the avalanche region.

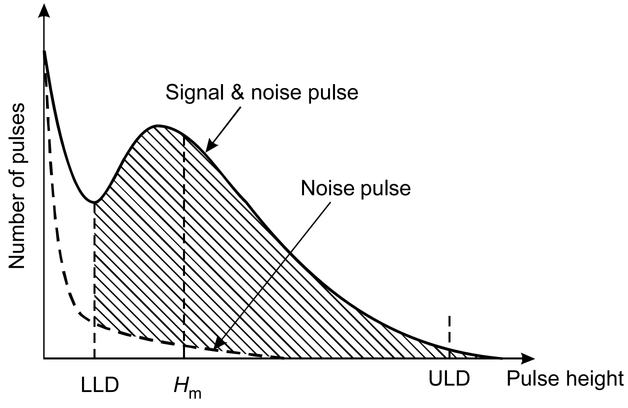


Figure A1.4.14. Typical PHD.

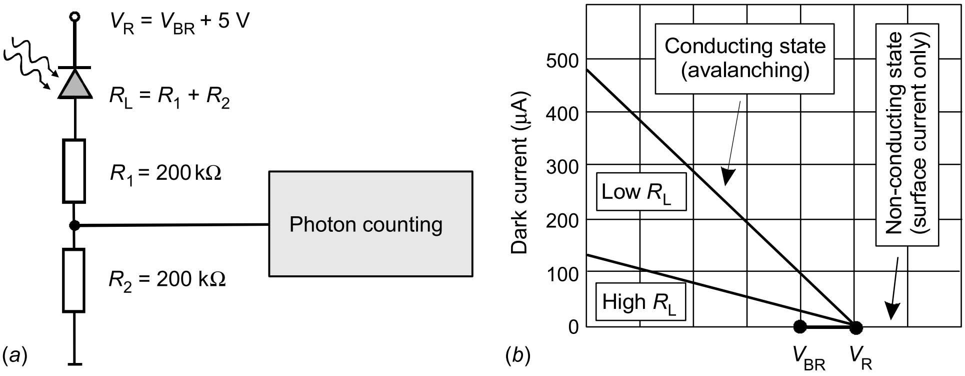


Figure A1.4.15. Passive quenched circuit (*a*) and the load-line of the circuit (*b*).

For the currents $(V_R - V_{BR})/R_L$ much greater than I_{latch}, the diode remains conducting. If the current $(V_R - V_{BR})/R_L$ is much less than I_{latch}, the diode switches almost immediately to the non-conducting state. When R_L is large, the photodiode is non-conducting, and the operating point is at $V_R - I_d R_L$ in the non-conducting state. Following an avalanche breakdown, the device recharges to the voltage $V_R - I_d R_L$ with the time constant CR_L where C is the total device capacitance including stray capacitance.

The rise-time is fast (e.g. 5–50 ns), decreases as $V_R - V_{BR}$ increases, and is very dependent on the capacitances of the load resistors, leads, etc. The jitter is typically the same order of magnitude as the rise-time.

To avoid an excessive dead-time when operating at a large voltage above V_{BR}, an 'actively quenched' circuit can be used. Active quenching can greatly increase this performance and has become commercially available over the last few years. Now, recharging can be very rapid through a small load resistor. Alternatively, the bias voltage can be maintained but the load resistor is replaced by a transistor that is kept off for a short time after an avalanche, and turned on for a period sufficient to recharge the photodiode. The response time is limited by the transistor switching rather than an *RC* circuit and has been reduced to as low as 50 ns, and the timing of photons can be made with resolution as high as 20 ps. In this mode, no amplifiers are necessary and single photon detection probabilities of up to approximately 50% are possible.

Photon-counting is also advantageous where gating coincidence techniques are employed for signal retrieval. Other applications in which APD operates in this mode are used in: lidar, astronomical observation, optical fibre test and fault location, optical range finding, ultrasensitive fluorescence, etc.

A1.4.2.5 High-speed photoreceivers

A variety of different applications and measurement techniques have been developed for high-speed detection. Most of them are described in section B2.5. Here we focus mainly on the speed limitation of the response of p–i–n junction and Schottky barrier—two types of photodiodes which have the highest speed response.

In general, in comparison with Schottky barriers, the p–n junction photodiodes indicate some important advantages. The thermionic emission process in a Schottky barrier is much more efficient than the diffusion process and therefore for a given built-in voltage, the saturation current in a Schottky diode is usually several orders of magnitude higher than in the p–n junction. In addition, the built-in voltage of a Schottky diode is smaller than that of a p–n junction with the same semiconductor.

However, high-frequency operation of p–n junction photodiodes is limited by the minority-carrier storage problem. In other words, the minimum time required to dissipate the carriers injected by the forward bias is dictated by the recombination lifetime. In a Schottky barrier, electrons are injected from the semiconductor into the metal under forward bias if the semiconductor is n-type. Next they thermalize very rapidly ($\approx 10^{-14}$ s) by carrier–carrier collisions, and this time is negligible compared to the minority-carrier recombination lifetime.

In p–i–n photodiode an undoped i-region (p^- or n^-, depending on the method of junction formation) is sandwiched between p^+ and n^+ regions. Because of the very low density of free carriers in the i-region and its high resistivity, any applied bias drops entirely across the i-region, which is fully depleted at zero bias or very low value of reverse bias.

The response speed of p–i–n photodiode is ultimately limited either:

- by transit time across i-region;
- or by circuit parameters (*RC* time constant).

Influence of diffusion time (τ_d) of carriers to the depletion region, which is inherently a relatively slow process, can be neglected since the generation of carriers occurs mainly in high-field i-region. If the photodetector is not fully depleted, this time is taken for collection of carriers generated outside the depletion layer. Carrier diffusion may result in a tail (a deviation between input and output signal) in the response time characteristics.

The transit time of the p–i–n photodiode is shorter than that obtained in a p–n photodiode even though the depletion region is longer than in the p–n photodiode case due to carriers travel at near their saturation velocity virtually the entire time they are in the depletion region (in p–n junction the electric field peaked at the p–n interface and then rapidly diminished).

The transit time of carriers across i-layer depends on its width and the carrier velocity. Usually, even for moderate reverse biases the carriers drift across the i-layer with saturation velocity. The transit time can be reduced by reducing the i-layer thickness. The fast photodiode must be thin to minimize the time that it takes for the photogenerated electrons and holes to traverse the depletion region. A thin photodiode has high capacitance per unit area and therefore must also be small in diameter to reduce the *RC* time constant.

The second component of the response time (t_r) is dependent mainly on the photodetector capacitance and input resistance of the preamplifier. The detector capacitance is a function of the area (A), the zero bias junction potential V_0, dielectric constant ($\varepsilon_0\varepsilon_s$), impurity concentrations of acceptors (N_a) and donors (N_d), and the bias voltage (V_b). If we assume that the external bias V_b is large compared to the V_0, and we have p^+–n abrupt junction, then the diode has a capacitance given by

$$C_d = \frac{A}{2}(2q\varepsilon_s\varepsilon_0 N_d)^{1/2} V_b^{-1/2}. \tag{A1.4.27}$$

In this case, lower *RC* time constant, and therefore, improved bandwidth can be achieved with the use of smaller detectors area and higher bias voltage. In practice, junctions are rarely abrupt; however, it still remains true that the capacitance decreases with increasing reverse bias.

The time constant is given by

$$\tau_{RC} = \frac{(R_L + R_s)R_{sh}}{R_L + R_s + R_{sh}} C = R_{eq} C \tag{A1.4.28}$$

where R_s is the series resistance of the photodiode, C is the sum of the photodiode and input preamplifier capacitances, R_{eq} is the equivalent resistance of the photodiode and load resistance.

The rise time dependent on time constant is described by the formula

$$t_r = 2R_{eq}C. \tag{A1.4.29}$$

The p–i–n photodiode has a 'controlled' depletion layer width, which can be tailored to meet the requirements of photoresponse and bandwidth. A tradeoff is necessary between response speed and quantum efficiency. For high response speed, the depletion layer width should be small but for high quantum efficiency (or responsivity) the width should be large.

Increases in bias voltage will usually increase carrier velocities and, therefore, reduce transit times but result in higher dark current and noise.

The detector response time is of the form

$$t_T = (t_r^2 + t_t^2 + t_d^2)^{1/2}. \tag{A1.4.30}$$

If we take equation (A1.4.29), the 3 dB cut-off frequency is given by

$$f_{-3\,dB} = (2\pi R_{eq}C)^{-1}. \tag{A1.4.31}$$

Usually, a designer can increase the bandwidth only by using smaller detectors and/or by reducing the amplifier's input resistance.

A good figure of merit for comparing sensitivity to high-speed signals of similar photodiodes is the *Response factor* × *Bandwidth product*. Response factor is the photodiode responsivity multiplied by the impedance seen by the photodiode. Assuming that two photodiodes have single pole high frequency roll-off, the one with the highest *Response factor* × *Bandwidth product* will provide the greatest response to an ultra fast pulse.

A1.4.2.6 Noise models of first stages of photoreceivers

A noise equivalent circuit of the first stage of a photoreceiver with a voltage type preamplifier is shown in figure A1.4.16. The signal current generator I_{ph} represents the detected signal. Noises in a detector (photodiode) are represented by three noise generators: I_{nph} is the shot noise originating from a photocurrent, I_{nd} is the shot noise of a dark current, I_{nb} is the shot noise from a background current. If the input resistance of a preamplifier is high, the value of the load resistance depends on the chosen bias resistor. The load (bias) resistor R_L affects both the level of the detector signal and its noise. The noise current generator I_{nR} is the thermal noise current and excess noise of the load resistance R_L. Since the

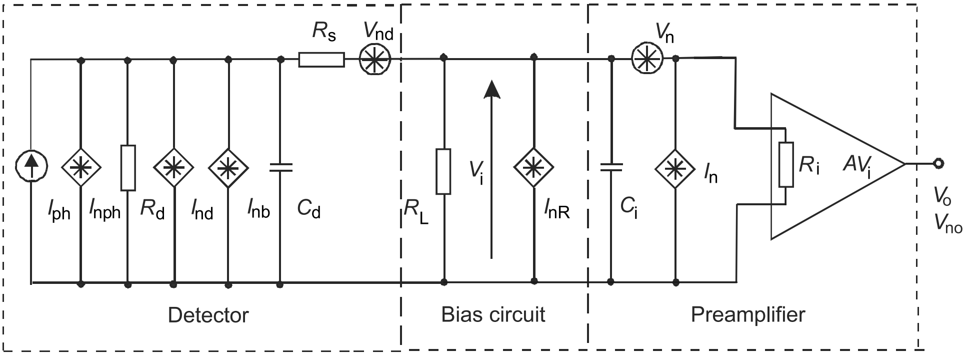


Figure A1.4.16. Equivalent scheme of photodetector–voltage type preamplifier circuit (C_d is the detector capacity, R_i and C_i are the input resistance and capacitance of the preamplifier, respectively).

thermal noise of I_{nR} is inversely related to the square root of the resistance, so R_{L} must be large. For the lowest noise system, at very low frequency, the detector will be the dominant noise source, but at higher frequencies amplifier noise becomes increasingly important.

Expression for signal-to-noise ratio at the first stage of a photoreceiver with voltage type preamplifier results from scheme A1.4.16:

$$\frac{S}{N}=\frac{I_{\mathrm{ph}}}{\left[I_{\mathrm{nph}}^2+I_{\mathrm{nd}}^2+I_{\mathrm{nb}}^2+I_{\mathrm{n}}^2+\frac{4kT\Delta f}{R_{\mathrm{L}}}+\left(\frac{V_{\mathrm{n}}}{R_{\mathrm{L}}}\right)^2\right]^{1/2}}. \tag{A1.4.32}$$

The numerator represents the photocurrent and the denominator represents the equivalent input noise of photodetector–preamplifier circuit. The first three components determine the noise originating from the photocurrent, the dark current and the background, the fourth component is the current noise of the preamplifier, the fifth is the thermal noise of the resistance R_{L} and the last one is the voltage noise contribution of the preamplifier.

The noise in electrical circuits is often a function of frequency. For high frequencies, the noise equivalent signal current, $I_{\mathrm{n\,total}}$, is given by

$$I_{\mathrm{n\,total}}^2=I_{\mathrm{nph}}^2+I_{\mathrm{nd}}^2+I_{\mathrm{nb}}^2+V_{\mathrm{nd}}^2\omega^2C_{\mathrm{d}}^2+I_{\mathrm{n}}^2+\frac{4kT\Delta f}{R_{\mathrm{L}}}+V_{\mathrm{n}}^2\omega^2(C_{\mathrm{d}}+C_{\mathrm{i}})^2 \tag{A1.4.33}$$

where V_{nd} is the voltage noise of the serial resistance, R_{s}.

The input capacitance of the preamplifier, C_{i}, may be considered to lie across the amplifier input on the photodiode side of the noise generator, V_{n}. Thus, with this assumption the noise generator, V_{n}, is not a true input noise generator as generally understood, as it should normally lie on the input side of all components (except, possibly, the noise current generator, which would have the same effect whichever side V_{n} it were to be connected to). The justification for this approach is firstly, that the amplifier capacitance is conveniently grouped with C_{d}, but secondly it enables V_{n} to be a 'white' noise generator.

Re-writing equation (A1.4.33) we obtain

$$I_{\mathrm{n\,total}}^2=\left[I_{\mathrm{nph}}^2+I_{\mathrm{nd}}^2+I_{\mathrm{nb}}^2+I_{\mathrm{n}}^2+\frac{4kT\Delta f}{R_{\mathrm{L}}}\right]+\omega^2[V_{\mathrm{nd}}^2C_{\mathrm{d}}^2+V_{\mathrm{n}}^2(C_{\mathrm{d}}+C_{\mathrm{i}})^2]. \tag{A1.4.34}$$

There are thus two terms, a white noise term in the first setoff square brackets and a second term that gives a noise current increasing in proportional to frequency. Although a capacitor does not add noise, the detector noise voltage (V_{nd}), and preamplifier noise voltage (V_{n}) is increased by the C_{d} and the $C_{\mathrm{d}}+C_{\mathrm{i}}$ respectively, as is evident from the coefficient of that term in equation A1.4.34. Analysing equation A1.4.34, we see that for matching an amplifier to a detector, it is important to minimize the sum of $I_{\mathrm{n}}+V_{\mathrm{n}}^2\omega^2(C_{\mathrm{d}}+C_{\mathrm{i}})^2$.

The sensitivity of an optical receiver is most conveniently expressed in terms of its NEP. This is defined as the optical power necessary to make the signal current, I_{ph} equal to the noise current $I_{\mathrm{n\,total}}$, i.e.

$$\mathrm{NEP}=(I_{\mathrm{n\,total}}^2)^{1/2}\frac{h\nu}{\eta q}. \tag{A1.4.35}$$

Many data sheets show the NEP figure for a detector, unfortunately. If we connected a preamplifier, the performance will often dependent on the amplifier noise source, critically combined with parasitic features of the photodiode, particularly its capacitance and parallel resistance (equation (A1.4.33)).

It is possible to achieve a high value of the signal-to-noise ratio at the first stage of a photoreceiver when using a voltage type preamplifier by using a high resistance value for R_L, and ensuring low current noise I_n, and low voltage noise V_n of the preamplifier. Of course, high resistance value of a R_L causes narrowing of a photoreceiver bandwidth.

Figure A1.4.17 presents the noise equivalent scheme of the first stage of a photoreceiver using a transimpedance preamplifier. In this circuit, noise sources of a detector are identical as for the case shown in figure A1.4.16, where R_{sh} is the shunt resistance of a detector. Preamplifier noise is represented by the voltage source V_n and the current source I_n. Thermal noises from a feedback resistor are represented by the current source I_{nf}.

From the arrangement in figure A1.4.17, it can be shown that the equivalent input noise is the square root of the sum of squares of noise components from: the photocurrent I_{ph}, the dark current of a detector I_d, the background current I_b, thermal noise of the resistor R_f, the current I_n, and the voltage V_n noise from a preamplifier. Thus, the signal-to-noise ratio is of the form

$$\frac{S}{N} = \frac{I_{ph}}{\left[I_{nph}^2 + I_{nd}^2 + I_{nb}^2 + I_n^2 + \frac{4kT\Delta f}{R_f} + \left(\frac{V_n}{R_f}\right)^2\right]^{1/2}}. \tag{A1.4.36}$$

For high frequencies, the last term of the denominator should contain parallel combination of all impedances across the input of the preamplifier, e.g. R_f, R_d, $(\omega C_d)^{-1}$, and $(\omega C_i)^{-1}$.

In a photoreceiver using transimpedance preamplifier, an identical bandwidth can be obtained by choosing a feedback resistance R_f much higher than the resistance $R = R_d \| R_L \| R_i$, which would be possible in a simpler photoreceiver with a voltage type preamplifier. Thus, comparing formulae (A1.4.32) and (A1.4.36), it can be noticed that for the same bandwidth the signal-to-noise ratio is higher for transimpedance preamplifier than for a classic one. In practice, it means that transimpedance amplifiers can have wider bandwidths yet retain the low noise characteristics of high-impedance preamplifiers.

In p–n and p–i–n photodiodes, the basic source of noise is shot noise originating from the photocurrent I_{ph}, the dark current I_d, and the background radiation current I_b. In these photodetectors, the thermal noises of detector resistance and noises of active elements of a preamplifier can also play a significant role.

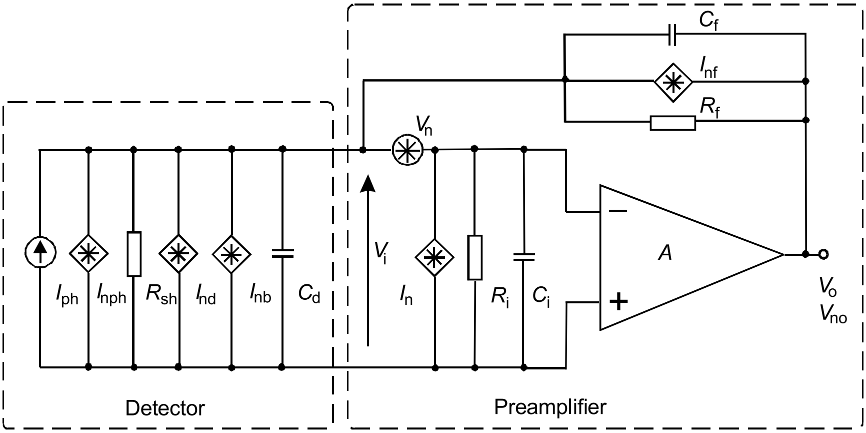


Figure A1.4.17. Equivalent scheme of the first stage of a photoreceiver with transimpedance preamplifier. Reproduced from [6].

All the preamplifier noises V_{n} and I_{n} can be substituted for one equivalent current noise:

$$I_{\mathrm{a}}^2 = \frac{1}{\Delta f}\int_0^{\Delta f}(I_{\mathrm{n}}^2 + V_{\mathrm{n}}^2|Y|^2)\mathrm{d}f \tag{A1.4.37}$$

where Y is the input admittance of an amplifier.

As stated earlier, photodiodes can operate with either voltage or transimpedance preamplifiers. For the equivalent schemes shown in figures A1.4.16 and A1.4.17, the signal-to-noise ratio is of the form:

$$\frac{S}{N} = \frac{I_{\mathrm{ph}}}{\left[2q(I_{\mathrm{ph}} + I_{\mathrm{d}} + I_{\mathrm{b}})\Delta f + \frac{4kT\Delta f}{R_{\mathrm{L}}} + I_{\mathrm{a}}^2\right]^{1/2}}. \tag{A1.4.38}$$

The first term represents the shot noise component of the photocurrent, the dark current and the background, whereas the second term is the thermal noise of load resistance of a photodetector, and the third term is the preamplifier noise. For the circuit shown in figure A1.4.17, the resistance R_{L} is the feedback resistor, R_{f}.

Let us consider a few special cases. Assuming that the signal current is higher than the dark current, I_{d} can be omitted in equation (A1.4.38). This is valid when the dark current is insignificant or if the received optical power is high. The assumption can also be made that the shot noise significantly exceeds thermal noise when the optical power has a high level. For low frequency FET transimpedance receivers, the shot noise limit is usually obtained when the output signal exceeds 50 mV. It means that the term $4kT\Delta f/R_{\mathrm{L}}$ can be omitted. If additionally $I_{\mathrm{a}}^2 \ll 2q\Delta f(I_{\mathrm{ph}} + I_{\mathrm{d}})$, the expression for signal-to-noise ratio can be simplified to the form

$$\frac{S}{N} = \left(\frac{I_{\mathrm{ph}}}{2q\Delta f}\right)^{1/2} = \left(\frac{\eta\Phi_{\mathrm{e}}\lambda}{2hc\Delta f}\right)^{1/2} = \left(\frac{\eta A E_{\mathrm{e}}\lambda}{2hc\Delta f}\right)^{1/2} \tag{A1.4.39}$$

where λ is the wavelength of incident radiation, Φ_{e} is the incident radiation flux (in W), and E_{e} is the detector's irradiance. A photoreceiver, the signal-to noise ratio of which is described by formula (A1.4.39), is limited only by a shot noise. This noise is also called quantum limited one—see equation (A1.4.12). Unfortunately, it is not always the case that the high optical power reaches a photoreceiver. If the power of an optical signal is low, the shot noise is negligible in relation to the thermal noise, then

$$\frac{S}{N} = \frac{\eta q \Phi_{\mathrm{e}}}{h\nu}\left(\frac{R_{\mathrm{L}}}{4kT\Delta f}\right)^{1/2}. \tag{A1.4.40}$$

It is evident that when a photoreceiver is limited by the thermal noise, it is thermally dependent (see equation (A1.4.13)).

When thermal noise is limited, it can be seen by analysing equation (A1.4.40) that the signal-to-noise ratio increases directly in proportion to the received optical power. Thus, in the range of a thermally dependent photoreceiver, small changes, e.g. of path transmission efficiency will cause significant differences in the signal-to-noise ratio of the received signal. In the quantum limited systems, an increase in the optical power by $\Delta\Phi_{\mathrm{e}}$ [dB] gives an improvement in the signal-to-noise ratio of only half the change $\Delta\Phi_{\mathrm{e}}$ when expressed in dB.

One way of overcoming the preamplifier noise is to use a detector having a high degree of internal gain, prior to connection to this device. There are two main types of detector that can be used; the photomultiplier, which is rather large and inconvenient, and the avalanche photodiode. The main advantages are apparent at high frequency, when amplifier noise contribution is more troublesome, because of stray capacitances lowering the impedance across the amplifier input.

Let us consider the signal-to-noise ratio in the first stage of photoreceiver with an avalanche photodiode. The high sensitivities of APDs can be obtained due to a phenomenon of avalanche multiplication, which significantly increases the current signal generated at the output of the detector and improves the signal-to-noise ratio. Of course it does not influence the noises from the load resistance and the amplifier noises, but increases the signal as usually only noise originating from the signal current and dark current (quantum noise) are dominant. Unfortunately, at high gain levels, the random mechanism of carrier multiplication (M) introduces excess noise, in the form of higher shot noise, which eventually exceeds the noise level resulting only from primary generation of unequilibrium carriers.

At moderate gain levels, the dominant source of the noise in an avalanche photodiode is the signal and dark current shot noises, which are multiplied. However, if the signal power is increased by a factor M^2, the noise power increases by a factor M^{2+x}. The factor x is typically between 0.3 and 0.5 for silicon APDs and between 0.7 and 1.0 for germanium and III–V alloy APDs.

Knowing the total noise, a signal-to-noise ratio can be determined as

$$\frac{S}{N} = \frac{MI_{\text{ph}}}{\left\{2q\Delta f\left[I_{\text{ph}}M^{2+x} + I_{\text{s}} + (I_{\text{b}} + I_{\text{db}})M^{2+x}\right] + \frac{4kT\Delta f}{R_{\text{L}}}F_{\text{n}}\right\}^{1/2}}. \tag{A1.4.41}$$

The numerator of this equation determines the photocurrent and denominator noises. The first term of the denominator is a shot noise term and the second one represents the thermal noise of load resistance with a preamplifier noise (F_{n} is the noise factor of preamplifier, I_{db} is the bulk leakage current component of primary dark current, and I_{s} is the surface leakage current of a dark current). Shot noise components except the surface leakage component of the dark current are multiplied, so the photocurrent, current from a background, and bulk leakage current components of the primary dark current are multiplied. When M is large, the thermal and amplifier noise term becomes insignificant and the S/R ratio decreases with increasing M. Therefore an optimum value of the multiplication factor M_{op} exists which maximizes the S/N ratio.

Figure A1.4.18 presents the typical dependence of signal current, thermal noise, shot noise and total noise on the avalanche multiplication factor [6]. For low values of multiplication factor, the signal

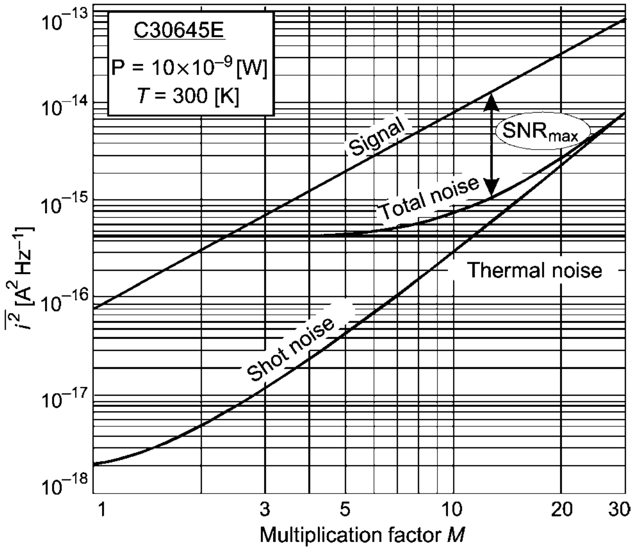


Figure A1.4.18. Dependence of signal, thermal noise, shot noise and total noise on avalanche multiplication factor.

amplitude is usually lower than the total noise amplitude. In this range, thermal noises are dominant at the low optical input levels usually used with APDs. For high multiplication values, thermal noises become less important but shot noises are then significant. There is a supply voltage for which a distance between the line representing the signal and the curve representing the total noise is the largest, corresponding to when the *S/N* ratio is maximum. Of course, these parameters will change with temperature, because the dark current of a photodiode and the avalanche multiplication factor are both strongly thermally dependent.

A1.4.3 Advanced method of signal detection

A1.4.3.1 Signal averaging

When measuring a small repetitive or steady-state signal in the presence of noise, we can often improve the results by making a number of measurements and taking their average. Figure A1.4.19 presents a scheme of signal detection system for steady-state signals with an analogue integrator.

If radiation from both a signal source and the background is incident on a detector surface, the desired signal will appear with noise at the detector output. It is amplified in a preamplifier and next reaches the integrator input. Let us assume that before beginning the measurement cycle, the voltages at inputs of both the differential amplifier of the integrator and the integrator output are all equal to zero. If a voltage V_i appears at the integrator input, a current I_1 will flow through the resistor R. This current causes changes the capacitor C, to give a voltage $V_o(T)$ at the system output, where:

$$V_o(T) = -\frac{1}{RC}\int_0^T V_i\,dt + V_o(0) \tag{A1.4.42}$$

where $V_o(0)$ is the initial voltage at the capacitor at $t = 0$ and $RC = \tau$ is the time constant of the integrator.

The voltage at the system output is inversely proportional to the time constant. In practice, the capacitor C is short-circuited at the beginning of a measuring cycle by means of the switch K. So, the second constant term of equation (A1.4.42) can be omitted. Opening the switch K initiates the measuring cycle, during which the signal integration occurs. If the assumption is taken that there is a constant voltage of the value v at the system input, then, after a time t, the output voltage will be directly proportional to the product vt. Thus, the signal-to-noise ratio at the integrator output takes a form

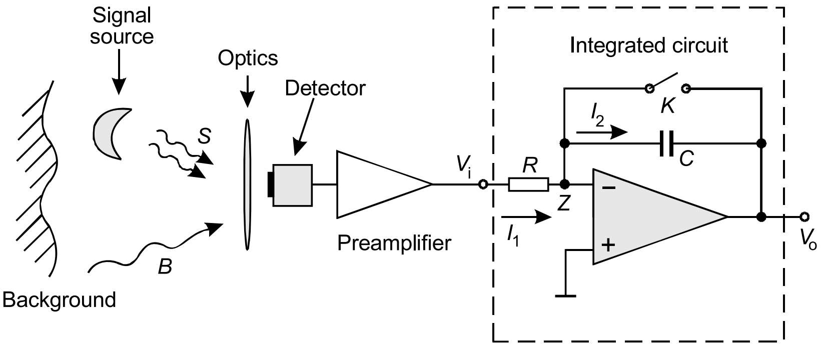


Figure A1.4.19. Analogue integrator used to collect detected signal level.

$$\frac{S}{N} = v\left(\frac{2t}{S_n}\right)^{1/2} \tag{A1.4.43}$$

where S_n is the input white noise power density. This equation tells that the signal-to-noise ratio can increase linearly with the square root of the integration time t (for $t \ll RC$).

The integrating process in mathematical form is equal taking a series of measurements and summing each of the received values. If we make N measurements, each integrated over the period t, and then the received results are added, the signal-to-noise ratio is described as

$$\frac{S}{N} = v\left(\frac{2Nt}{S_n}\right)^{1/2}. \tag{A1.4.44}$$

In this case, signal-to-noise ratio increases proportionally to square root of a number of the performed measurements (or the total measurement time Nt). Similarly as in equation (A1.4.43), the time constant of an integrator does not affect the value of a signal-to-noise ratio. In real conditions, the value of the time constant should be chosen to ensure the desired level of the output signal after an appropriate integration time t [7].

One disadvantage of the simple integrator is that $1/f$ noise, electronic offsets, and background light level charges can all degrade the performance. We shall now describe another signal recovery method that avoids these problems.

A1.4.3.2 Lock-in amplifier

A lock-in amplifier uses phase-sensitive detection to improve the signal-to-noise ratio in cw experiments. For phase-sensitive detection, the analogue signal should be modulated at some reference frequency. This enables the noise component to be filtered out, even when it may initially be many times stronger than the signal itself. If the detected signal of interest is modulated by a carrier signal of defined phase and frequency, it is possible to separate this signal from the noise component. This requires the availability of a low-noise, stable reference frequency ω_{ref} of the same frequency as the carrier signal. In optical measurements, it is usually achieved by initially modulating a stable light source with this reference frequency ω_{ref}. The light transmitted, reflected or scattered from the test sample, now heavily attenuated but still modulated with ω_{ref}, is then measured. Measurement information is provided by the amplitude of the signal that falls on the detector (figure A1.4.20). The reference signal is connected to the

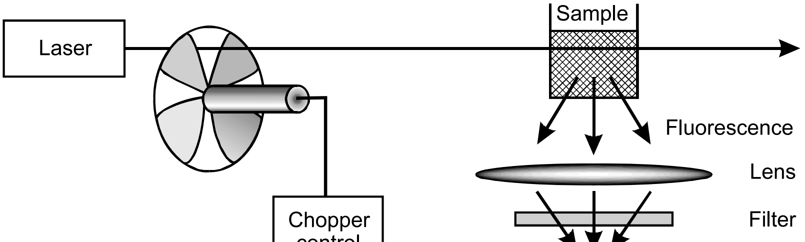


Figure A1.4.20. Use of a lock-in amplifier in fluorescence detection.

lock-in amplifier via the reference input (figure A1.4.21). The first stage of the lock-in amplifier is usually a programmable ac amplifier which matches the signal amplitude to a suitable level, ensuring it does not overload (saturate or 'clip') the following electronic components.

In the diagram below, the reference signal is a square wave of frequency ω_{ref}. This might be the clock output from a function generator. If the sine output from the function generator is used to excite the experiment, the response might be the signal waveform shown below.

In order to simplify the analysis, we shall assume that the system is noise-free.

The signal at the optical detector is

$$V_{\mathrm{d}} = V_{\mathrm{S}} \sin(\omega t + \phi) + V_{\mathrm{n}} \tag{A1.4.45}$$

where V_{S} is the signal amplitude and ϕ is the signal phase.

The bandpass filter, centred at ω_{ref}, removes all noise frequencies which are outside a defined bandwidth around ω_{ref}. This is to give some pre-filtering, to stop the following circuits being so easily overloaded by high levels of noise. The next stage is the phase sensitive demodulator (PSD). Here, the filtered signal is multiplied by a square wave signal of the same frequency ω_{ref} and, ideally, the same phase as the signal at this point.

The internal reference is

$$V_{\mathrm{H}} = V_{\mathrm{L}} \sin(\omega_{\mathrm{L}} t + \phi_{\mathrm{ref}}). \tag{A1.4.46}$$

The output of the PSD is simply the product of two sine waves

$$\begin{aligned} V_{\mathrm{psd}} &= V_{\mathrm{S}} V_{\mathrm{L}} \sin(\omega t + \phi)\sin\left(\omega_{\mathrm{L}} t + \phi_{\mathrm{ref}}\right) \\ &= \frac{1}{2} V_{\mathrm{S}} V_{\mathrm{L}} \cos\left[(\omega - \omega_{\mathrm{L}})t + \phi - \phi_{\mathrm{ref}}\right] - \frac{1}{2} V_{\mathrm{S}} V_{\mathrm{L}} \cos\left[(\omega + \omega_{\mathrm{L}})t + \phi + \phi_{\mathrm{ref}}\right]. \end{aligned} \tag{A1.4.47}$$

The PSD output is two ac signals, one at the difference frequency $(\omega - \omega_{\mathrm{L}})$ and other at the sum frequency $(\omega + \omega_{\mathrm{L}})$. The phase difference ϕ between the square wave and the signal is important here. If both signals are exactly in phase, the resulting output signal reaches a maximum (figure A1.4.22). If this is not the case, the desired ac component will not be so well detected in the subsequent processes and will actually give zero output if the phase difference is 90°. For this reason, lock-in amplifiers are equipped with a phase shifter which is used to bring the two signals into phase, either manually or automatically.

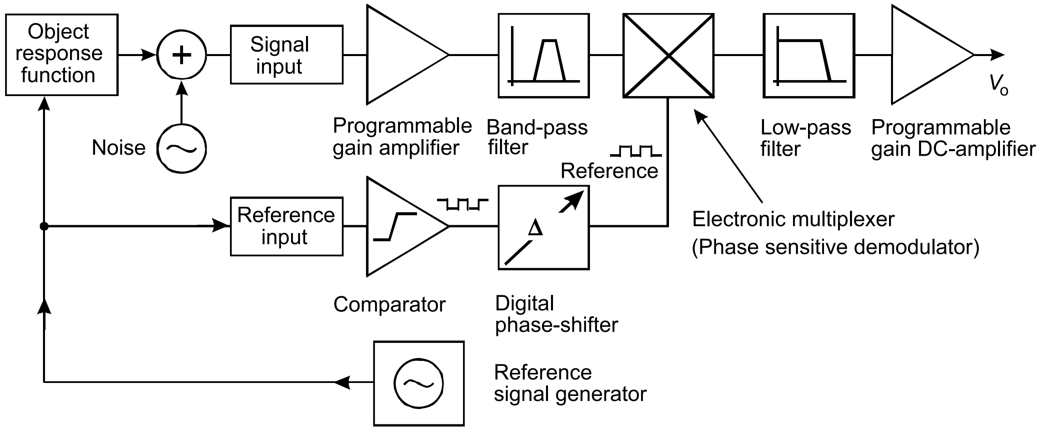

Object response function
+
Noise
Signal input
Programmable gain amplifier
Band-pass filter
Reference
Low-pass filter
Programmable gain DC-amplifier
V_o
Electronic multiplexer (Phase sensitive demodulator)
Reference input
Comparator
Δ
Digital phase-shifter
Reference signal generator

Figure A1.4.21. Schematic diagram of a lock-in amplifier. Reproduced from [8].

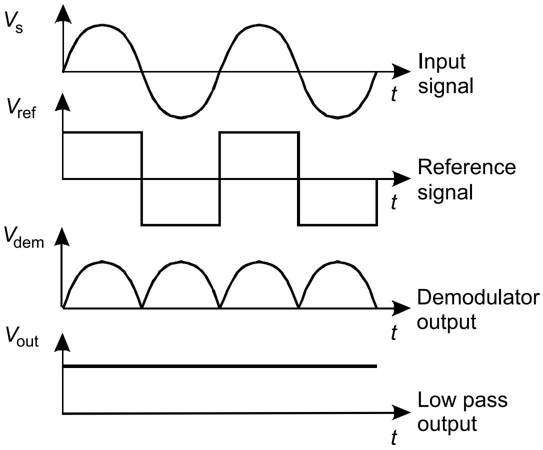


Figure A1.4.22. Signal processing where input signal and reference signal are in phase.

Sometimes these lock-ins are used with the references set 90° apart, or 'in-phase' and 'quadrature' signals are obtained. Then the desired signal is the square root of the sum of the squares of in-phase and quadrature components, and budding is not possible. The resulting signal includes contribution from the sum and difference frequencies of the two components.

A low pass filter is then used to remove any ac components from the dc signal. The lower the cut-off frequency of the filter, the better the suppression of the unwanted noise components. However, this also leads to a longer filter time constant and therefore a longer measurement time. Any changes in optical signal level that occur faster than the filter time constant can, of course, no longer be observed. The low pass filter effectively removes any high frequency noise remaining following the demodulator process. However, if ω_{ref} equals ω_{L}, the difference frequency component will be a dc signal. In this case, the filtered PSD output will be

$$V_{\mathrm{out}} = \frac{1}{2} V_{\mathrm{S}} V_{\mathrm{L}} \cos(\phi - \phi_{\mathrm{ref}}). \tag{A1.4.48}$$

The dc signal is proportional to the signal amplitude.

The final stage of the lock-in consists of a programmable dc amplifier which further amplifies the smoothed signal. At the output, the result is ideally a noise-free dc signal whose amplitude is directly proportional to the strength of ac signal at the lock-in input. A further advantage of the lock-in amplifier is that it removes '$1/f$' noise components, as only noise at frequencies close to the modulation frequency is effectively observed after the electronic processing.

In the above considerations, we have assumed that the chopped signal and the reference output share the same phase. This may not always be the case.

A1.4.3.3 Boxcar detection systems

Phase-sensitive detection systems are used for measurements of steady periodic signals or ones that have a relatively slow varying level. Frequently, it is desired to measure the amplitude envelope of periodical signals. For this purpose, various measuring methods can be used. One of these methods is the so-called 'boxcar' detection system, which performs synchronous integration (figure A1.4.23). This method allows measurement of periodical signals of complex shapes, even when the signal amplitude is lower than the

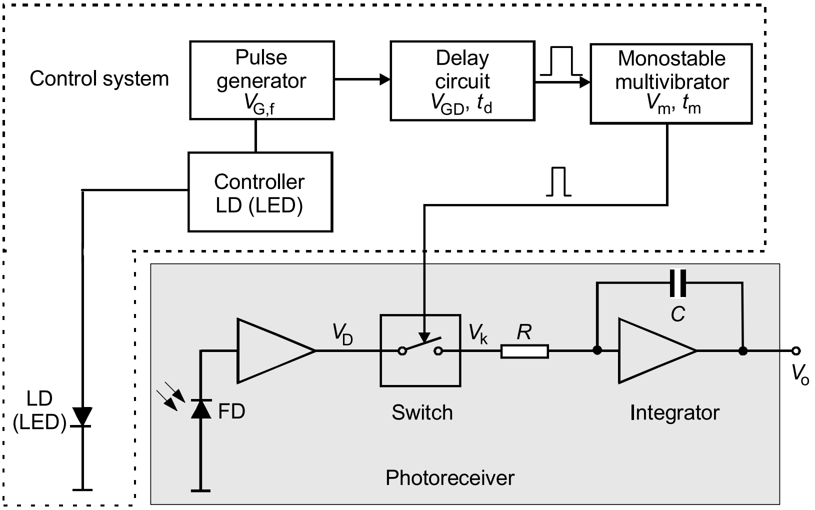

Control system
Pulse generator $V_{G,f}$
Delay circuit V_{GD}, t_d
Monostable multivibrator V_m, t_m
Controller LD (LED)
LD (LED)
FD
V_D
V_k
R
C
V_o
Switch
Integrator
Photoreceiver

Figure A1.4.23. Analogue synchronous integration system.

level of the first stage noise of a photoreceiver. To apply this method, two basic conditions must be fulfilled:

- the measured signal should be periodic and repeatable;
- a trigger signal should be available, which can be used to tell the measurement system when each signal cycle begins. (This latter signal could be derived using a phase-lock-loop if the periodic signal is available for a long enough time.)

The detection system consists of a signal source, detector–preamplifier system, switch, integrator and a control system. The control system includes: a clock, delay circuit and a multivibrator. Pulses from the clock generator are used to drive simultaneously an optical signal source and a delay circuit in the signal processor. The output signal from the delay circuit triggers the multivibrator, which is turn closes a switch. The switch on time of the key switch depends on the pulse duration of the multivibrator. A semiconductor laser (LD) or electro-luminescence diode (LED) is usually used as the optical source.

We shall assume that the control generator operates at a frequency $f = 1/T$. The delay time of the clock pulse controlling initiation of the multivibrator is t_d and the key-switch on time is t_m, as determined by the mono-stable multivibrator. The leading edge of each clock pulse starts the light pulse generation. Figure A1.4.24 shows the temporal variation courses of voltages at critical points of the boxcar detection system. A signal from the detector output is sent after amplification to an electronic switch and then to an integrator to average the signal.

The signal $V_k(t)$ reaching the integrator is described by the formula

$$\begin{aligned} V_k(t) &\equiv V_D(t) \qquad \text{for } t_d \le t \le t_d + t_m \\ V_k(t) &\equiv 0 \qquad \text{for other } t. \end{aligned} \tag{A1.4.49}$$

We shall now calculate the voltage at the output of the integrator. We assume that the voltage at the integrator output is initially equal to zero and that we make measurements for m periods of an input waveform. In order to simplify the initial analysis, we shall assume that the system is noise free. For this

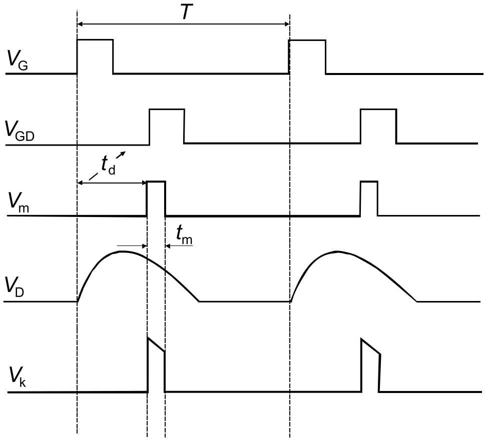


Figure A1.4.24. Control and data waveforms in phase-sensitive detection systems.

assumption, the voltage at the system output is

$$V_o(t_d) = -\frac{1}{RC} m \int_0^T V_k(t)\mathrm{d}t = -\frac{1}{RC} m \int_{t_d}^{t_d+t_m} V_D(t)\mathrm{d}t \qquad \text{(A1.4.50)}$$

where R and C determine the resistance and the capacitance values of the integrator, respectively. The minus sign in this equation is present because the voltage signal is applied to an inverting input. If the time interval is sufficiently small, the signal level will not change significantly between the times, t_d and $t_d + t_m$. Thus, formula (A1.4.50) takes the form

$$V_o(t_d) = -\frac{m}{RC} V_D(t) t_m. \qquad \text{(A1.4.51)}$$

It results from this formula that the voltage value at the output of the integrator $V_o(t)$ is directly proportional to the instantaneous value of the voltage at the preamplifier output $V_D(t)$, signal cycles m, and pulse duration t_m, but is inversely proportional to the time constant of the integrator. The boxcar detection system exhibits an increase in the amplitude of the output signal in proportion to the number of measuring cycles m.

The signal-to-noise ratio of the boxcar detection system if a white noise input voltage source is now considered to be present is given by

$$\frac{S}{N} = \frac{V_D(t)}{V_{ni}} m(2\Delta f t_m)^{1/2}. \qquad \text{(A1.4.52)}$$

An increase in the signal-to-noise ratio is obtained with a larger number of measuring cycles and longer time of key switch on, the improvement being in proportion to m.

The above measuring system is called a boxcar detection system, because the samples are added synchronously with the measuring cycles of a signal, like loading wagons (or boxcars) on a train. Boxcar detection systems are very effective in recovering information from repetitive signals when the noise level is quite high. Improvement in the signal-to-noise ratio, obtained due to signal measurement for m cycles is, of course, at the expense of greater measurement time, being given by mT.

A drawback of the method is that the integrator is disconnected from the input for most of the time. A voltage measurement at the photodetector–preamplifier system is performed only during the time interval t_m/T (i.e. when the key is switched on). So, to measure a pulse shape, we have to repeat the measurement process up to T/t_m times for each t_d value. The time required to measure a pulse shape will be mT^2/t_m. The total measurement time can be reduced by increasing the time of a single sampling t_m. However, it constrains observation of the pulse details.

The boxcar detection system is not efficient because most of the signal power is ignored during the key switch off. This drawback can be avoided by using a parallel processing scheme, involving a multiplexed array of boxcar detection systems. To achieve this, an analogue multiplexer can be used to separate into parallel channels or, more frequently with modern technology, a system with digital path of signal processing is applied. Many digital oscilloscopes are now equipped with such parallel averaging schemes.

A1.4.3.4 Coherent detection

So far, we have considered systems that were based on the modulation of the light intensity in a transmitter and direct detection of this in a photoreceiver. It was not essential for the modulated light wave to be a coherent wave and its spectrum could be wide. These systems are simple and cheap but they have constraints on their transmission possibilities. The photoreceivers decoded only the information connected with the intensity or with the square of the electromagnetic field amplitude whereas information can be carried also by its phase and frequency. The possible photoreceiver sensitivity results from the basic noise limits, as the noise of the photodetector, preamplifier, and background.

To improve the signal-to-noise ratio, it would be an advantage to increase the photocurrent at the detector output. We have already noted that using photoreceivers with APDs had constraints resulting from additional multiplication noises. To avoid problems with direct detection, coherent detection, a method of receiving based on interference of two beams of coherent laser radiation, can be used [9–11]. Figure A1.4.25 illustrates the differences between coherent detection and direct detection.

Figure A1.4.25(*a*) presents a direct detection system. To narrow the received optical bandwidth, we have applied a filter to limit the spectral range of radiation reaching the detector—figure A1.4.25(*b*). For example, optical filters based on the Bragg effect can have 5-nm bandwidth for 1.56 μm wavelengths, corresponding to a detection band, Δf, equal to 600 GHz [12]. This wide bandwidth, of what is a reasonably narrow optical filter, illustrates that application of an optical filter cannot easily ensure narrow wavelength selection of a photoreceiver. A widely spaced Fabry–Perot filter can achieve bandwidth of several hundred MHz, but is costly and difficult to stabilize, so impractical for most systems. However, the simple addition of a beam-splitter and an additional coherent light source, a so-called, local oscillator, provides a coherent detection scheme, that can use either heterodyne and homodyne detection systems (see figure A1.4.25(*c*)).

Heterodyne detection technique has been commonly used for many years in commercial and domestic radioreceivers and also for the microwave range of an electromagnetic spectrum. The main virtues of this method of detection are higher sensitivity, higher and more easily obtained selectivity, plus the possibility of detection of all types of modulation and easier tuning over wide range [13]. Coherent optical detection has been developed since 1962, but compact and stable production of this system is more difficult, and the system is more expensive and troublesome than its radio-technique equivalent. The basic block diagram of heterodyne optical receiver is shown in figure A1.4.26.

Laser radiation containing information, is after being passed through an input optical filter and beam splitter, arranged to coherently combine or ‘mix’ with a light beam of a local oscillator at the detector surface. A beam-splitter can be made in many ways, the simplest being a glass plate with adequate refraction coefficient. In a general case, a device fulfilling such a role is called a direction

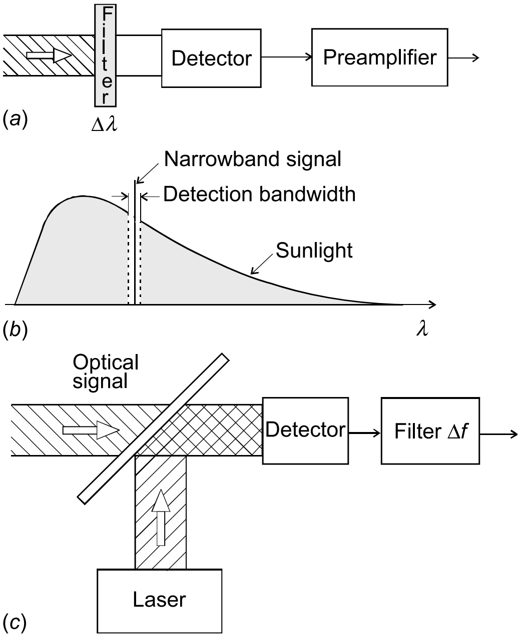

Filter
Detector
Preamplifier
(a)
Δλ
Narrowband signal
Detection bandwidth
Sunlight
(b)
λ
Optical signal
Detector
Filter Δf
Laser
(c)

Figure A1.4.25. Comparison of coherent versus incoherent optical detection.

coupler, as an analogy to microwave or radio devices. A detector used for signal mixing has to have a square-law characteristic to detecting electronic field of the light, but this is conveniently typical of most optical detectors (photodiode, photoconductor, photomultiplier, APD, etc). This signal is next amplified. An electrical filter of intermediate frequency (IF) extracts the desired difference component of the signal, which next undergoes a demodulation process. The design and operation principle of the subsequent electrical detector depends on the nature of the modulation of a signal. The signal from a load resistance passes through an output filter to a receiver output and by means of a local oscillator frequency controller it controls a laser. A frequency control loop is used for the local oscillator laser to maintain a constant frequency difference $\omega_L - \omega_s = \omega_p$ with the input signal. An indispensable condition

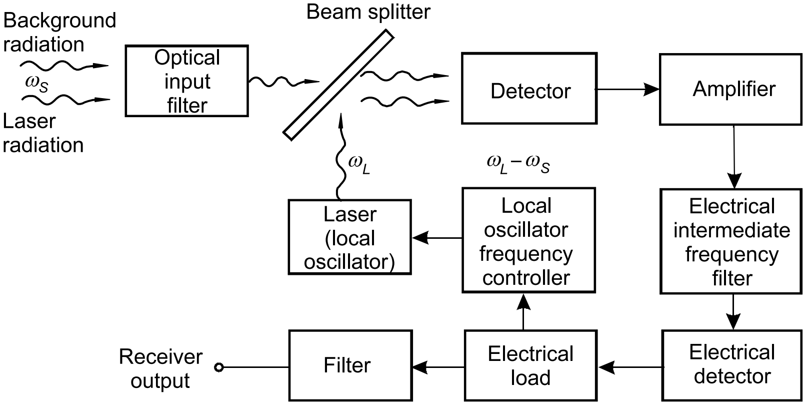

Background radiation
ω_S
Laser radiation
Optical input filter
Beam splitter
Detector
Amplifier
ω_L
$\omega_L - \omega_S$
Laser (local oscillator)
Local oscillator frequency controller
Electrical intermediate frequency filter
Receiver output
Filter
Electrical load
Electrical detector

Figure A1.4.26. Block diagram of heterodyne detection optical receiver.

for efficient coherent detection is to match the polarisation, and the shape of both waveforms of both beams to match the profile of the detector surface.

Expression for the signal-to-noise ratio at the heterodyne detection system with APD for $P_L \gg P_S$ is given by

$$\frac{S}{N} = \frac{\sqrt{2}R_i M\sqrt{P_S P_L}}{\left(2q\Delta f M^{2+x} R_i P_L + \frac{4kT\Delta fF}{R_L}\right)^{1/2}} \approx \left(\frac{R_i P_S}{q\Delta f M}\right)^{1/2}. \tag{A1.4.53}$$

Assuming a photodiode responsivity $R_i = \eta q/h\nu$, we have

$$\frac{S}{N} = \frac{\eta P_S}{h\nu\Delta f M^x} = 2\left(\frac{S}{N}\right)_{\text{quanta}}. \tag{A1.4.54}$$

Figure A1.4.27 shows a comparison of coherent detection sensitivity (solid lines) with the sensitivity of direct p–i–n photodiode detection ($M = 1$) for the same values of signal-to-noise ratio. Significant improvement in sensitivity can be observed for weak signals. Higher sensitivity of a detector ensures qualitatively better detection as increased information bit rate can permit longer communications links to be used between each regenerator circuit. In long-distance fibre telecommunications, however, the use of optical fibre amplifier has taken much of the impetus from development of the coherent receiver, although the latter still has the unique advantage of highly selective narrowband detection.

In heterodyne detection, the spectrum of laser modulation was shifted into an IF range, so selectivity of a photoreceiver depends on the bandwidth of the IF amplifier. This is arranged

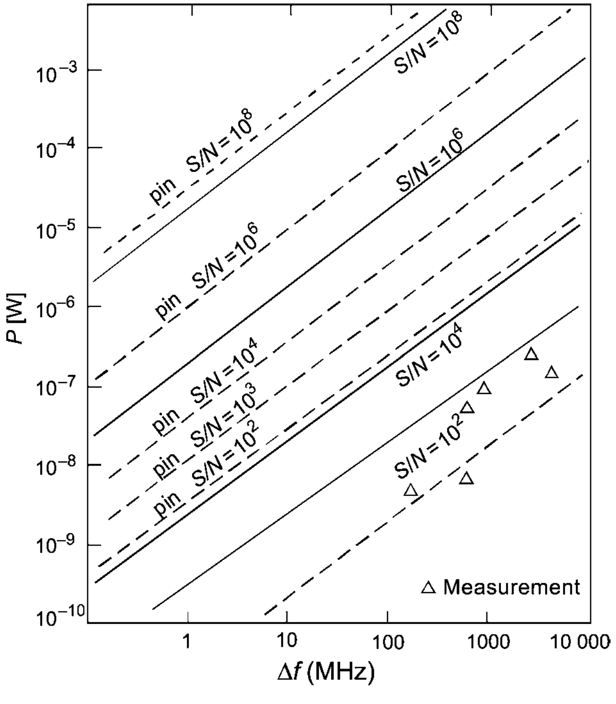


Figure A1.4.27. Sensitivity of the coherent photoreceiver (dashed line) and p–i–n photodiode (solid line). Reproduced from [14].

electronically, so it can easily be sufficiently narrow. Having narrow IF circuit bandwidth is especially important for detection of multichannel signals.

In practice, the technique of heterodyne detection is used for construction of Doppler velocimeters and laser rangefinders and as well as in spectroscopy (particular LIDAR systems). It may yet find application in more telecommunications systems.

If a signal frequency is equal to the frequency of a local oscillator, the IF frequency equals zero. It is a special case of coherent detection, so-called, homodyne detection.

In a homodyne detection optical receiver (figure A1.4.28), the incoming laser carrier is again combined with a reference wave from a local laser on a photodiode surface, but in this case both frequencies are the same. It does not contain two blocks, filter of IF frequency and demodulator which were in the heterodyne receiver.

The photodetector current in a homodyne receiver is given by:

$$I_{\mathrm{hom}} = R_{\mathrm{i}}M(P_{\mathrm{S}} + P_{\mathrm{L}}) + 2R_{\mathrm{i}}M(P_{\mathrm{S}}P_{\mathrm{L}})^{1/2}\cos\phi_{\mathrm{p}}(t). \tag{A1.4.55}$$

The first component is a direct-current component but the second one contains the useful information regarding the optical signal. The current at the detector output increases with increase in local oscillator power and with the optical receiver responsivity.

If the local oscillator power is high, the shot noise originating from a signal current, thermal noise and dark-current noise can be omitted. For amplitude modulation we have

$$\frac{S}{N} = \frac{(2R_{\mathrm{i}}M(P_{\mathrm{S}}P_{\mathrm{L}})^{1/2})^2 R_{\mathrm{L}}}{(2q\Delta f M^{2+x}R_{\mathrm{i}}P_{\mathrm{L}} + 4kT\Delta f F/R_{\mathrm{L}})R_{\mathrm{L}}} \approx \frac{2R_{\mathrm{i}}P_{\mathrm{S}}}{q\Delta f M^{x}} = \frac{2\eta P_{\mathrm{S}}}{h\nu\Delta f M^{x}}. \tag{A1.4.56}$$

As can be seen, the signal-to-noise ratio for homodyne detection is twice as high as heterodyne detection. This is basically because the homodyne detector allows direct addition or subtraction of the electrical fields, depending on whether the signal and local oscillator are in phase or 180° out of phase. With heterodyne detection, the relative phases change linearly with time, and mixing of signals is not effective when signals have 90 or 270° phase difference.

As it results from equation A1.4.55, homodyne detection derives the baseband modulation signal carrying the information directly. Thus, further electronic demodulation is not required.

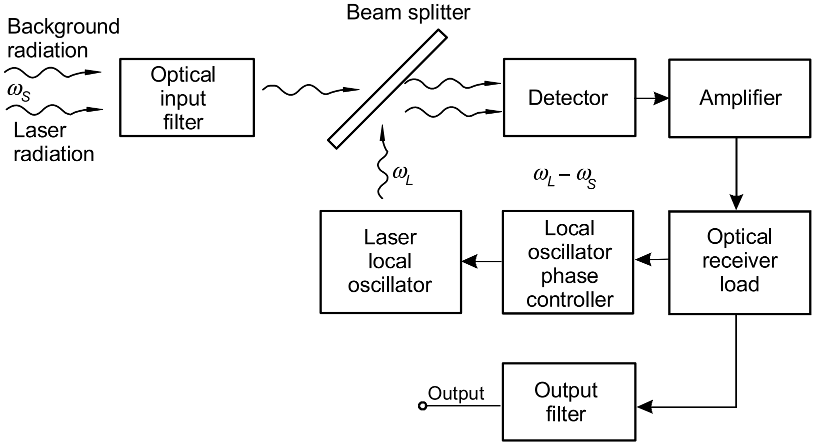


Figure A1.4.28. Block diagram of homodyne detection optical receiver.

Homodyne receivers are used in the most sensitive coherent systems. In practice, construction of such receivers is difficult because

- the local oscillator must be locked to keep a constant zero phase difference to the incoming optical signal, and this requires excellent spectral purity;
- power fluctuation of the local laser must be eliminated.

Constant difference of both laser phases can be achieved using an optical phase-locked loop.

The requirements for spectrum purity are less critical in the diversity systems [10, 12] in which the cosine component current expressed by equation A1.4.55 and also the quadrature current component proportional to $\sin \phi_p t$ are produced. A sum of vectors of these currents makes it possible to avoid influence of ϕ_p phase change, but as with the heterodyne system, the SNR suffers when phases are not identical.

It has been assumed that the local laser has no amplitude noise. In practice this type of noise is often the limiting factor, because the local laser power is strong compared with signal component $2(P_S P_L)^{1/2}$. If detecting these output waves by means of photodiodes gives the corresponding currents

$$I_1 \propto a^2 P_S + b^2 P_L + 2ab(P_S P_L)^{1/2} \cos[\omega_S t - \omega_L t + \phi_S(t) - \phi_L(t) + \pi/2] \tag{A1.4.57}$$

$$I_2 \propto b^2 P_S + a^2 P_L - 2ab(P_S P_L)^{1/2} \cos[\omega_S t - \omega_L t + \phi_S(t) - \phi_L(t) + \pi/2]. \tag{A1.4.58}$$

This noise is contained in the terms $b^2 P_L$ and $a^2 P_L$, respectively. If we assume a symmetrical beam combiner ($a = b$), and the currents I_1 and I_2 are subtracted, then the terms containing P_S and P_L cancel out, and so do their amplitude fluctuations. Due to the opposite sign of the third term in equations (A1.4.57) and (A1.4.58), the output signal from the subtractor is doubled. Using the two outputs in this way produces a balanced mixing receiver, using a beam splitter or fibre coupler (figure A1.4.29).

In order to make the interference of the signal wave and the local oscillator wave that is received more efficient, their polarization states must coincide. Due to random vibration in the fibre and temperature changes, mechanical strain in the fibre introduces birefringence, which changes with time. As a consequence, the polarization state of the signal received changes randomly.

The problems caused by a polarization mismatch can be overcome in the following ways by:

- using a polarization state controller;
- polarization scrambling (the polarization state is deliberately changed at the transmitting end);

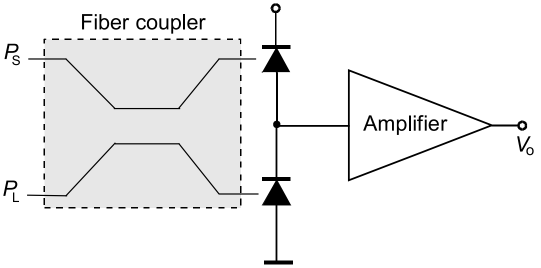


Figure A1.4.29. Balanced mixing receiver with fibre coupler and series connection of the photodiodes.

- use of polarization-maintaining fibres (this solution is more expensive);
- using polarization diversity (both the local optical wave and signal wave received are split into two orthogonal polarization states).

A1.4.4 Arrays of detectors and detectors for focal plane arrays

Many materials have been investigated to fabricate photodetectors [15–17]. Figure A1.4.30 shows the quantum efficiency of some of the detector materials used to fabricate arrays of ultraviolet (UV), visible and infrared detectors. AlGaN detectors are being developed in the UV region. Silicon p–i–n diodes are shown with and without antireflection coating. Lead salts (PbS and PbSe) have intermediate quantum efficiencies, while PtSi Schottky barrier types and quantum well infrared photodetectors (QWIPs) have low values. InSb can respond from the near UV out to 5.5 μm at 80 K. A suitable detector material for near-IR (1.0–1.7 μm) spectral range is InGaAs lattice matched to the InP. Various HgCdTe alloys, in both photovoltaic and photoconductive configurations, cover from 0.7 to over 20 μm. Impurity-doped (Sb, As, and Ga) silicon impurity-blocked conduction (IBC) detectors operating at 10 K have a spectral response cut-off in the range from 16 to 30 μm. Impurity-doped Ge detectors can extend the response out to 100–200 μm.

The term FPA refers to an assemblage of individual detector picture elements (‘pixels’) located at the focal plane of an imaging system. Although the definition could include one-dimensional (‘linear’) arrays as well as 2D arrays, it is frequently applied to the latter (see figure B8.14(*b*)). Usually, the optics part of an optoelectronic images device is limited only to focusing of the image onto the detectors array. These so-called ‘staring arrays’ are scanned electronically usually using circuits integrated with the arrays. The architecture of detector-readout assemblies has assumed a number of forms that are discussed below. The types of readout integrated circuits (ROICs) include the function of pixel deselecting, antiblooming on each pixel, subframe imaging, output preamplifiers, and may include yet other functions. Infrared imaging systems which use two-dimensional arrays belong to so-called ‘second generation’ systems.

Development in detector FPA technology has revolutionized many kinds of imaging in the past 25 years [18]. From γ rays to the infrared and even radio waves, the rate at which images can be acquired

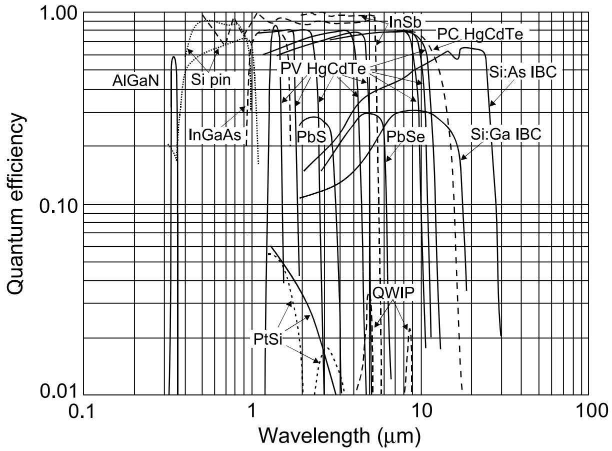


Figure A1.4.30. Quantum efficiency of UV, visible, and infrared detector arrays.

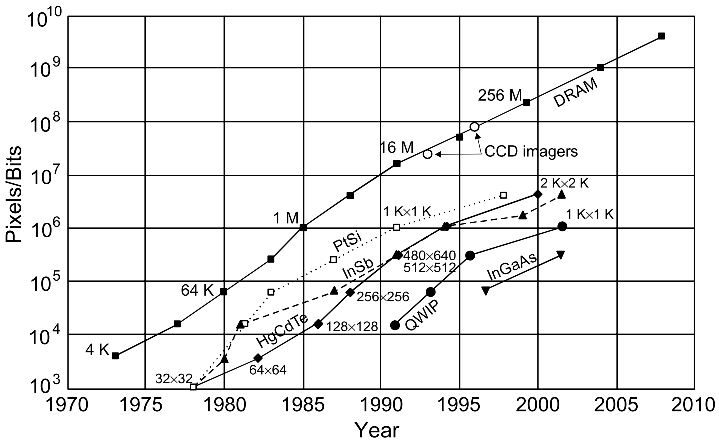


Figure A1.4.31. Increase in array format size over the past 30 years. Reproduced from [18].

has increased by more than a factor of a million in many cases. Figure A1.4.31 illustrates the trend in array size over the past 30 years. Imaging FPAs have developed in proportion to the ability of silicon ICs technology to read and process the array signals, and with ability to display the resulting image. FPAs have nominally the same growth rate as dynamic random access memory (DRAM) ICs (which have had a doubling-rate period of approximately 18 months; it is a consequence of Moore's Law, which predicts the ability to double transistor integration on each IC about every 18 months) but lag behind in size by about 5–10 years. ROICs are somewhat analogous to DRAM-only readouts, but require a minimum of three transistors per pixel, compared to one for each memory cell. Readouts are also analogous in terms of an emphasis on low noise inputs and generally maximum charge storage capacity. Charge coupled devices (CCDs) with close to 100 M pixels offer the largest formats. PtSi, InSb and HgCdTe have been following the pace of DRAM. In the infrared, 4 M pixel arrays are now available for astronomy applications.

A1.4.4.1 Monolithic arrays

In general, the architectures of FPAs may be classified as monolithic and hybrid. When the detector material is either silicon or a silicon derivative (e.g. platinum silicide, PtSi), the detector and ROIC can be built on a single wafer. There are a few obvious advantages to this structure, principally in the simplicity and lower cost associated with a directly integrated structure. Common examples of these FPAs in the visible and near infrared (0.7–1.0 μm) are found in camcorders and digital cameras. Two generic types of silicon technology provide the bulk of devices in these markets: CCDs and complementary metal–oxide–semiconductor (CMOS) imagers. CCD technology has achieved the highest pixel counts or largest formats with numbers approaching 10^8 (figure A1.4.32). CMOS imagers are also rapidly moving to large formats and are expected to compete with CCDs for the large format applications within a few years.

An example CCD array made by Dalsa is illustrated in figure A1.4.33. This array has approximately 25 M pixels, and can operate at 2.5 frames per second with 8 outputs. Although monolithic structures have been proposed for a wide variety of infrared detector materials over the past 30 years, only a few have been demonstrated. These include PtSi, and more recently PbS, PbTe, and uncooled microbolometers. Uncooled detector arrays are revolutionizing the infrared imaging community, by eliminating the need for expensive and high-maintenance coolers that are required by

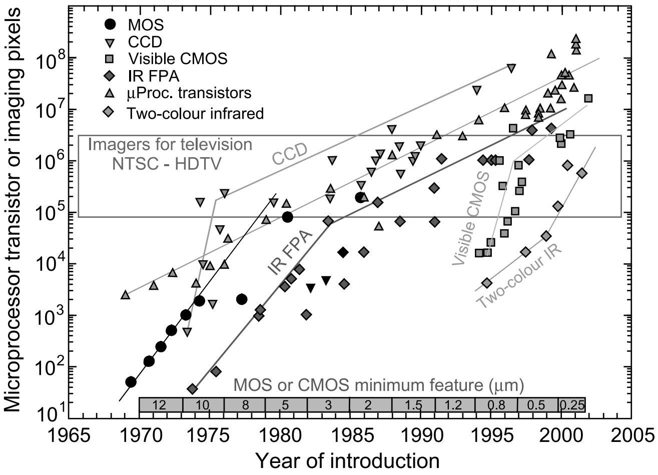


Figure A1.4.32. Imaging array formats compared with the complexity of microprocessor technology as indicated by transistor count. The timeline design rule of MOS/CMOS features is shown at the bottom. Reproduced from [18].

traditional infrared detectors. At present they can be built in formats as large as 480 × 640 array with $25 \times 25\,\mu m^2$ pixels.

CCD technology is very mature in respect to both the fabrication yield and the attainment of near-theoretical sensitivity. Figure A1.4.34 shows the schematic circuit for a typical CCD imager. The monolithic array is based on a metal–insulator–semiconductor (MIS) structure. Incident radiation generates electron–hole pairs in the depletion region of the MIS structure. The photogenerated carriers are first integrated in an electronic well at the pixel and subsequently transferred to slow and fast CCD shift registers. At the end of the CCD register, a charge carrying information on the received signal can be readout and converted into a useful signal. More information about CCD operation can be found in section B2.

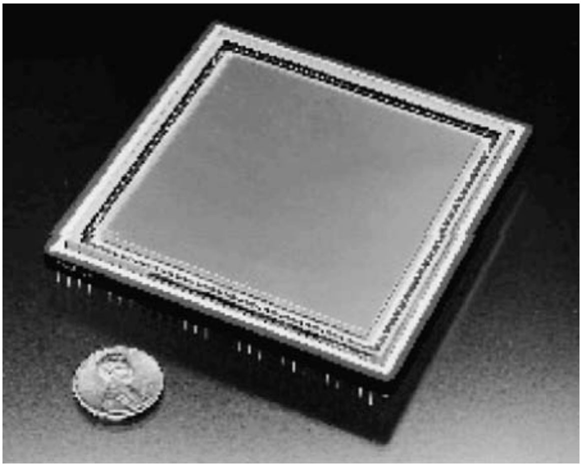


Figure A1.4.33. Silicon CCD visible array with 5040 × 5040 pixels. Pixel size is $12 \times 12\,\mu m^2$. The chip size is $60 \times 60\,mm^2$. Quantum efficiency is >20% at 900 nm. Reproduced from [19].

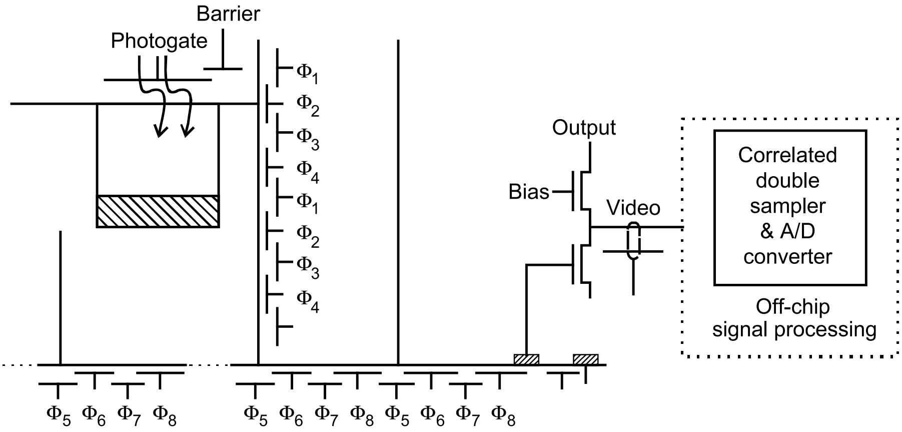

Barrier
Photogate
Φ1
Φ2
Φ3
Φ4
Φ1
Φ2
Φ3
Φ4
Output
Bias
Video
Correlated double sampler & A/D converter
Off-chip signal processing
Φ5 Φ6 Φ7 Φ8
Φ5 Φ6 Φ7 Φ8 Φ5 Φ6 Φ7 Φ8

Figure A1.4.34. Architecture of typical CCD image. Reproduced from [20].

At present, the following readout techniques are used in CCD devices:

- floating diffusion amplifier in each pixel;
- system with correlated double sampling (CDS);
- floating gate amplifier.

The floating diffusion amplifier, a typical CCD output preamplifier, can be implemented in each unit cell as shown in dotted box in figure A1.4.35. The unit cell consists of three transistors and the detector. Photocurrent is integrated onto the stray capacitance, which is the combined capacitance presented by the gate of the source follower T2, the interconnection, and the detector capacitance. The capacitance is reset to the voltage level V_R by supplying the reset clock (Φ_R) between successive integration frames. Integration of the signal charge makes the potential of the source follower input

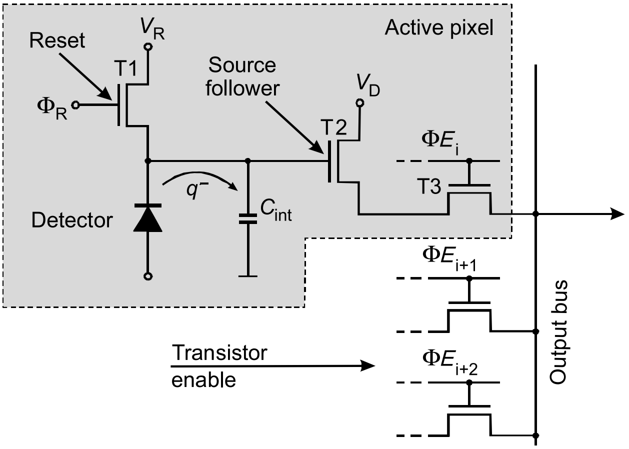

Reset
VR
T1
ΦR
Source follower
VD
T2
Active pixel
ΦEi
T3
q−
Cint
Detector
ΦEi+1
Transistor enable
ΦEi+2
Output bus

Figure A1.4.35. Unit cell with floating diffusion amplifier. Reproduced from [10].

node lower. The source follower is active only when the transistor T3 is clocked. The drain current of the source follower T2 flows through the enable transistor T3 and load resistor outside the array.

Figure A1.4.36 shows a preamplifier, in this example the source follower per detector (SFD), the output of which is connected to a clamp circuit. The output signal is initially sampled across the clamp capacitor during the onset of photon integration (after the detector is reset). The action of the clamp switch and capacitor subtracts any initial offset voltage from the output waveform. Because the initial sample is made before significant photon charge has been integrated, by charging the capacitor, the final integrated photon signal swing is unaltered. However, any offset voltage or drift present at the beginning of integration is, by the action of the circuit, subtracted from the final value. This process of sampling each pixel twice, once at the beginning of the frame and again at the end, and providing the difference is called CDS.

The value of the initial CDS sample represents dc offsets, low frequency drift and $1/f$ noise, and high-frequency noise; this initial value is subtracted from the final value, which also includes dc offset, low-frequency drift, and high-frequency noise. Since the two samples occur within a short period of time, the dc and lower-frequency drift components of each sample do not change significantly; hence, these terms cancel in the subtraction process.

The floating gate amplifier configuration is shown in figure A1.4.37. It consists of two MOSFET transistors, the source follower T2 and the zeroing transistor T1. The floating gate (reading gate) is in the same row as the CCD transfer gates. If a moving charge is under the gate, it causes a change in the gate potential of the transistor of the gate T2. At the preamplifier output, a voltage signal appears. This manner of readout does not cause degradation or decay of a moving charge hence the charge can be detected at many places. An amplifier, in which the same charge is sampled with several floating gates is called a floating diffusion amplifier.

The configuration of CCD devices requires specialized processing, unlike CMOS imagers which can be built on fabrication lines designed for commercial microprocessors. CMOS have the advantage that existing foundries intended for application specific integrated circuits (ASICs), can be readily used by adapting their design rules. Design rules of 0.18 μm are in production, with pre-production runs of 0.13 μm design rules already underway. As a result of such fine design rules, more functionality has been put into the unit cells of multiplexers and smaller unit cells, leading to large array sizes. Figure A1.4.32 shows the timelines for minimum circuit features and the resulting CCD, IR FPA and CMOS visible imager sizes with respect to imaging pixels. Along the horizontal axis is also a scale depicting the general availability of various MOS and CMOS processes. The ongoing migration to even finer lithographies

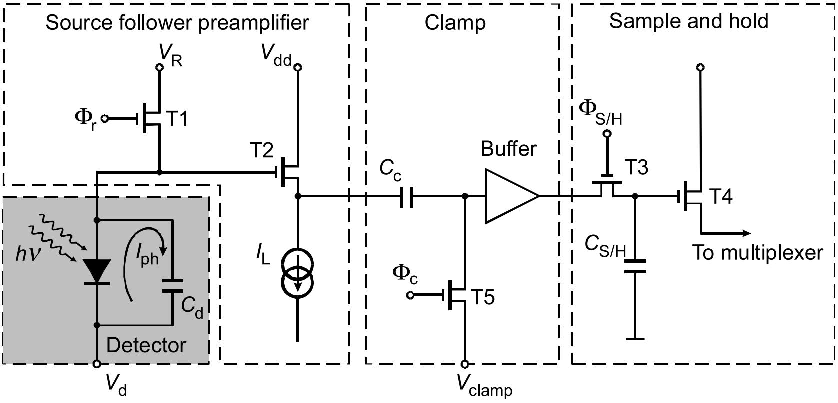


Figure A1.4.36. CDS circuit. Reproduced after [3].

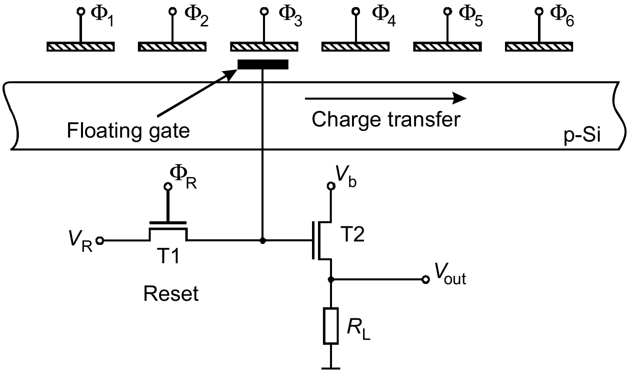

Φ_1
Φ_2
Φ_3
Φ_4
Φ_5
Φ_6
Floating gate
Charge transfer
p-Si
Φ_R
V_b
V_R
T1
Reset
T2
V_{out}
R_L

Figure A1.4.37. Floating gain amplifier circuit. Reproduced after [21].

will thus enable the rapid development of CMOS-based imagers having even higher resolution, better image quality, higher levels of integration and lower overall imaging system cost than CCD-based solutions. At present, CMOS having minimum features of ≤0.5 μm is also enabling monolithic visible CMOS imagers, because the denser photolithography allows low-noise signal extraction and high performance detection with the optical fill factor within each pixel. The silicon wafer production infrastructure which has put personal computers into many homes is now enabling CMOS-based imaging in consumer products such as digital still and video cameras.

A typical CMOS multiplexer architecture (figure A1.4.38) consists of fast (column) and slow (row) shift registers at the edges of the active area, and pixels are addressed one by one through the selection of a slow register, while the fast register scans through a column, and so on. Each photodiode is connected in parallel to a storage capacitor located in the unit cell. A column of diodes and storage capacitors is selected one at a time by a digital horizontal scan register and a row bus is selected by the vertical scan register. Therefore, each pixel can be individually addressed.

CMOS-based imagers use active or passive pixels [22–24] as shown, in simplified form, in figure A1.4.39. In comparison with passive pixel sensors (PPSs), active pixel sensors (APSs) apart from read

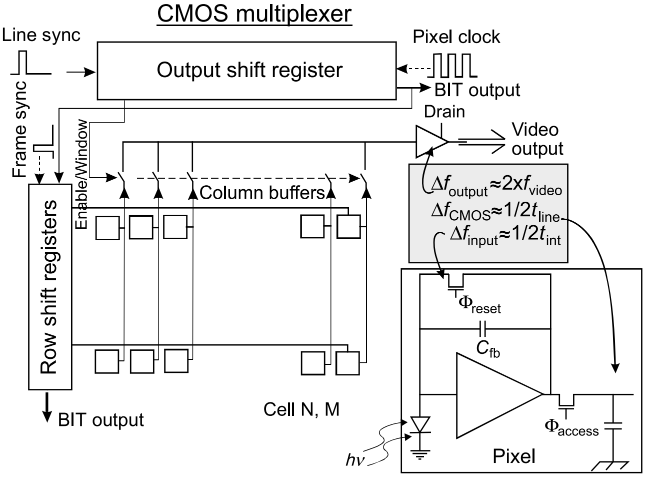

CMOS multiplexer
Line sync
Pixel clock
Output shift register
BIT output
Drain
Video output
Frame sync
Enable/Window
Column buffers
$\Delta f_{output} \approx 2x f_{video}$
$\Delta f_{CMOS} \approx 1/2t_{line}$
$\Delta f_{input} \approx 1/2t_{int}$
Row shift registers
Φ_{reset}
C_{fb}
Φ_{access}
BIT output
Cell N, M
$h\nu$
Pixel

Figure A1.4.38. CMOS multiplexing readout with CTIA detector interface. Reproduced from [22].

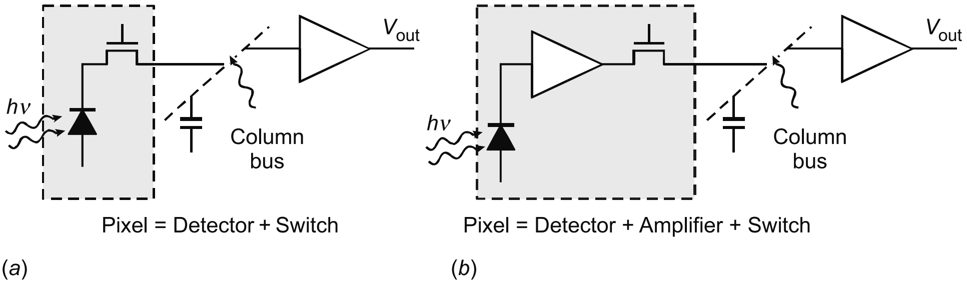

hv
Column bus
V_{out}
Pixel = Detector + Switch
(a)
hv
Column bus
V_{out}
Pixel = Detector + Amplifier + Switch
(b)

Figure A1.4.39. Passive (*a*) and active (*b*) pixel sensor. Reproduced from [22].

functions exploit some form of amplification at each pixel. PPSs have simple pixels consisting of as few as two components (a photodiode and a MOSFET switch). As a result, circuit overhead is low and the optical collection efficiency (fill factor (FF)) is high even for monolithic devices. A large optical FF of up to 80% maximizes signal selection and minimizes fabrication cost by obviating the need for microlenses. Microlenses, typically used in CCD and CMOS APS imagers for visible application, concentrate the incoming light into the photosensitive region when they are accurately deposited over each pixel (figure A1.4.40). When the FF is low and microlenses are not used, the light falling elsewhere is either lost or, in some cases, creates artefacts in the imagery by generating electrical currents in the active circuitry.

APSs incorporate transistors in each pixel to convert the photogenerated charge to a voltage, amplify the signal voltage and reduce noise. Adding these components, however, reduces the FF of monolithic imagers to about 30–50% in 0.5 μm processes at a 5–6 μm pixel pitch or in 0.25 μm processes at a 3.3–4.0 μm pixel pitch [22].

A1.4.4.2 Hybrid arrays

Ultraviolet and infrared imagers are most commonly built with a hybrid structure. Visible hybrids have also been built for specific applications. Hybrid FPAs detectors and multiplexers are fabricated on different substrates and mated with each other by flip-chip bonding or loophole interconnection (figure B8.17). In this case, we can optimize the detector material and multiplexer independently. Indium bump bonding of readout electronics, first demonstrated in the mid-1970s, provides for multiplexing the signals from thousands of pixels onto a few output lines, greatly simplifying the interface between the sensor and the system electronics.

Key to the development of ROICs has been the evolution in input preamplifier technology. This evolution has been driven by increased performance requirements and silicon processing technology improvements. A brief discussion of the various circuits is given below.

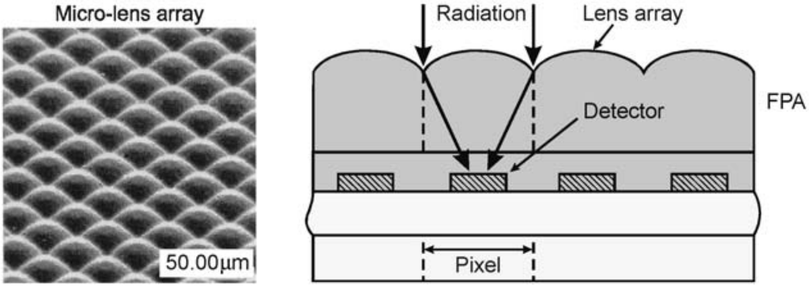


Figure A1.4.40. Micrograph and cross-sectional drawing of microlensed hybrid FPAs. Reproduced from [22].

The direct injection (DI) circuit was one of the first integrated readout preamplifiers and has been used as an input to CCDs and visible imagers for many years. This readout configuration requires the area at the unit cell (<20 × 20 μm) to be minimized. Photon current in DI circuits is injected, via the source of the input transistor, onto an integration capacitor (figure A1.4.41). As the photon current (figure A1.4.41(*b*)) charges the capacitor throughout the frame a simple charge integration takes place. Next a multiplexer reads out the final value and the capacitor voltage is reset prior to the beginning of the frame. To reduce detector noise, it is important that a uniform, near-zero-voltage bias be maintained across all the detectors.

Feedback enhanced direct injection (FEDI) is similar to direct injection except that inverting amplifier is provided between the detector node and the input MOSFET gate (figure A1.4.41 dashed line). The inverting gain provides feedback to yield better control over the detector bias at different photocurrent levels. It can maintain a constant detector bias at medium and high backgrounds. The amplifier reduces the input impedance of the DI and, therefore, increases the injection efficiency and bandwidth. The minimum operating photon flux range of the FEDI is an order of magnitude below that of the DI, thus the response is linear over a larger range than the DI circuit.

The combined source follower/detector (SFD) unit cell is shown in figure A1.4.42. The unit cell consists of an integration capacitance, a reset transistor (T1) operated as a switch, the source-follower transistor (T2), and selection transistor (T3). The integration capacitance may just be the detector capacitance and transistor T2 input capacitance. The integration capacitance is reset to a reference voltage (V_R) by pulsing the reset transistor. The photocurrent is then integrated on the capacitance during the integration period. The ramping input voltage of the SFD is buffered by the source follower and then multiplexed, via the T3 switch, to a common bus prior to the video output buffer. After the multiplexer read cycle, the input node is reset and the integration cycle begins again. The switch must have very low current leakage characteristics when in the open state, or this will add to the photocurrent signal. The dynamic range of the SFD is limited by the current voltage characteristics of the detector. As the signal is integrated, the detector bias changes with time and incident light level. The SFD has low noise for low bandwidth applications such as astronomy and still has acceptable signal-to-noise at very low backgrounds (e.g. a few photons per pixel per 100 ms). It is nonlinear at medium and high backgrounds, resulting in a limited dynamic range. The gain is set by the detector responsivity and the combined detector plus source–follower–input capacitance. The major noise sources are the kTC noise (resulting from resetting the detector), MOSFET channel thermal and MOSFET $1/f$ noise.

The capacitor–feedback transimpedance amplifier (CTIA) is a reset integrator and addresses broad range of detector interface and performance requirements across many applications. The CTIA consists of an inverting amplifier with a gain of A, the integration capacitance C_f placed in a feedback loop, and

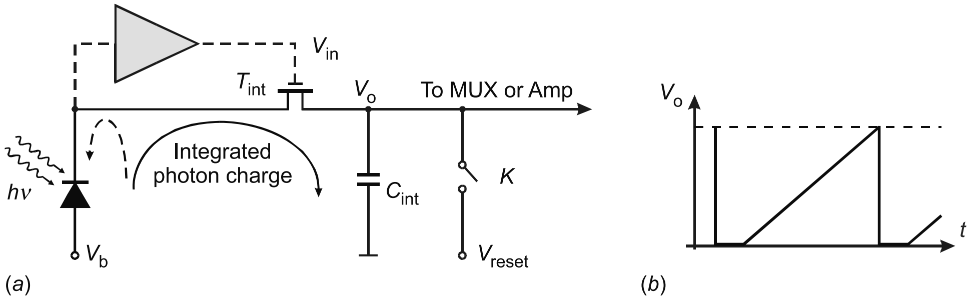


Figure A1.4.41. Direct injection readout circuit. Reproduced from [25].

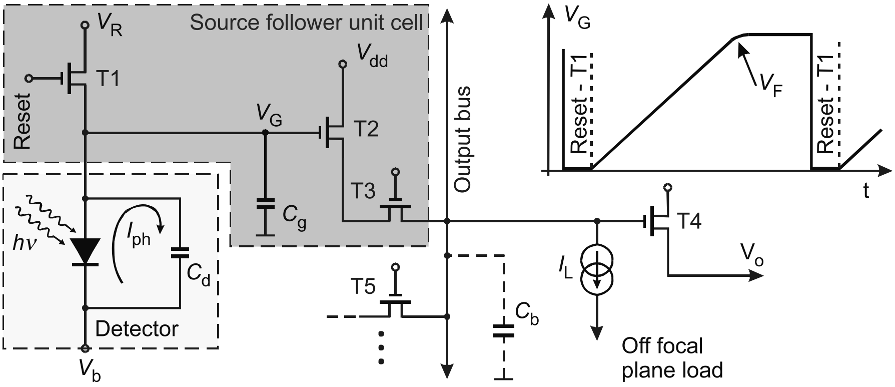


Figure A1.4.42. Schematic of source-follower per detector unit cell.

the reset switch *K* (figure A1.4.43). The photoelectron charge causes a slight change in a voltage at the inverting input node of an amplifier. The amplifier responds with a sharp reduction in output voltage. As the detector current accumulates over the 'frame time', uniform illumination results in a linear ramp at the output. At the end of integration, the output voltage is sampled and multiplexed to the output bus. Since the input impedance of the amplifier is low, the integration capacitance can be made extremely small, yielding low noise performance. The feedback, or integration, capacitor sets the gain. The switch *K* is cyclically closed to achieve reset. The CTIA provides low input impedance, stable detector bias, high gain, high frequency response and a high photon current injection efficiency. It has very low noise from low to high backgrounds.

The resistor load (RL) gate modulation circuit is shown in figure A1.4.44. It was introduced to extend the SFD performance advantages to high-irradiance backgrounds and dark currents. This circuit uses the photocurrent to modulate the gate voltage and thereby induce an output current in the MOSFET. The drain current of the MOSFET transistor accumulates onto an integration capacitor. In high background irradiance, this circuit provides a design that can reject much of these background components, as when the background alone is present on the detector, the bias on the detector or the load resistor can be adjusted to give negligible drain current or integration of a charge. When the signal is then applied, the transistor drain current increases with photon current and thereby allows some level

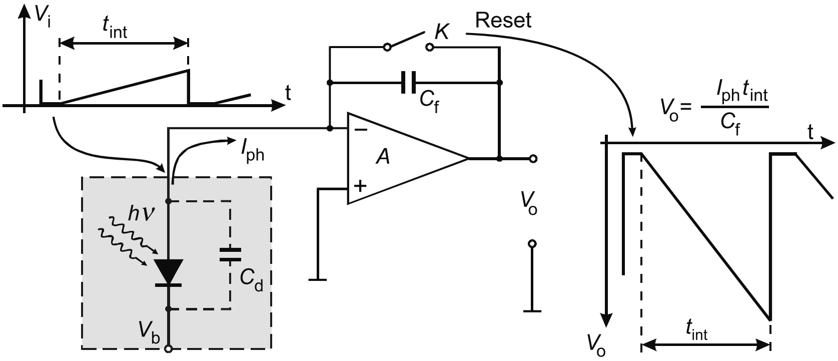


Figure A1.4.43. Schematic of a capacitive transimpedance amplifier unit cell.

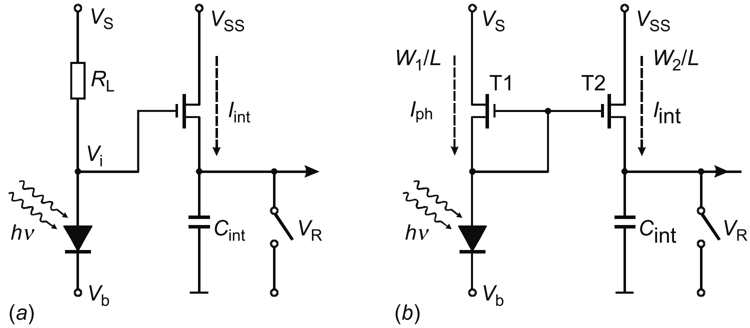


Figure A1.4.44. Resistor load gate modulation (*a*) and current mirror gate modulation (*b*).

of background flux rejection. The load resistor is designed such that it has low $1/f$ noise, excellent temperature stability, and good cell to cell uniformity.

The current mirror (CM) gate modulation, (figure A1.4.44(*b*)), extends readouts to very high background levels. In this current mirror preamplifier, the MOSFET replaces the resistor of the RL circuit. The photon current flowing into the drain of the first of two closely matched transistors includes a common gate to source voltage change in both transistors. This results in a similar current in the second transistor. If the source voltage, V_s and V_{ss}, of the two matched transistors are connected, both will have the same gate to source voltages which will induce a current in the output transistor identical to the detector current flowing through the input transistor. In this circuit, the integration current is a linear function of a detector current. This CM interfaces easily to direct access or CCD multiplexers and has low area requirements for the unit cell. The CM circuit requires gain and offset corrections for most applications. The advantages over the RL circuit include its better linearity and absence of a load resistor.

A wide variety of detector materials has been adapted to the hybrid format. UV, visible, and infrared arrays most commonly employ a photodiode structure. Photodiodes are preferred to photoconductors because of their relatively high impedance, which matches directly into the high input impedance stage of an FET readout circuit and also allows lower power dissipation. Mesa photodiodes are used in AlGaN, InSb, and HgCdTe detectors, whereas planar photodiodes are used in Si, PtSi, Ge, HgCdTe, InGaAs, and InSb detectors. A third photodiode structure—used exclusively with HgCdTe detectors—is the high-density vertically-integrated photodiode, or loophole photodiode [16].

An alternative hybrid detector for the long wavelength IR region (8–14 μm) is the quantum well infrared photoconductor (QWIP). These high impedance detectors are built from alternating thin layers (superlattices) of GaAs and AlGaAs. A distinct feature of n-type QWIPs is that the optical absorption strength is proportional to the electric-field polarization component of an incident photon in a direction normal to the plane of the quantum wells. For imaging, it is necessary to couple light uniformly to two-dimensional arrays of these detectors, so a diffraction grating is incorporated on one side of the detectors to redirect a normally incident photon into propagation angles more favourable for absorption.

Also extrinsic silicon detectors can form high impedance photoconductors in a hybrid configuration and be operated out to about 30 μm. Shallow, hydrogen-like impurities, such as phosphorus, antimony, or arsenic, provide electrons which can be ionized with photon energies in the range of 30–50 meV, depending upon the dopant and concentration used. Arrays of QWIP as well as extrinsic silicon detectors in a 1024 × 1024 format have been demonstrated [18].

The largest hybrid arrays have been principally built for astronomy where dark currents as low as 0.02 electrons per second are measured at 30 K. The most recent development is the 2 × 2 K format of

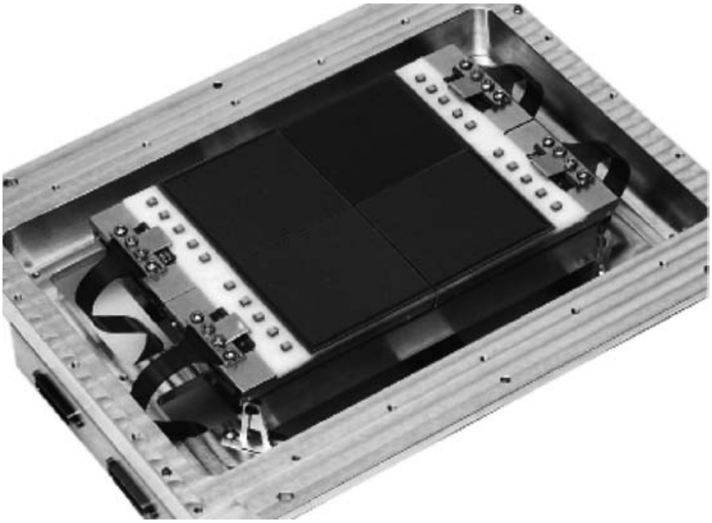


Figure A1.4.45. The three-side buttable arrays can be arranged in $2 \times n$ array configurations to build up large rectangular formats [26].

InSb (1–5 μm, 25 μm pixels) and HgCdTe (1–3 μm, 18 μm pixels). Individual arrays can be arranged in groups to give larger format configurations. Two-side buttable 2052 × 2052 arrays can be arranged in a 4104 × 4104 format, as illustrated in figure A1.4.45. The three-side buttable arrays can be arranged in $2 \times n$ array configurations to build up large rectangular formats [26].

In the infrared spectral region, third generation systems are now being developed. In this class of detector, two main competitors, HgCdTe photodiodes and QWIPs are considered. Those that provide enhanced capabilities like larger number of pixels, higher frame rates, better thermal resolution as well as multicolour functionality and other on-chip functions are considered as third generation IR systems. Multicolour capabilities are highly desirable for advanced IR systems. Systems that gather data in separate IR spectral bands can discriminate both absolute temperature and unique signatures of objects in the scene. By providing this new dimension of contrast, multiband detection also offers advanced colour processing algorithms to further improve sensitivity compared to that of single-colour devices.

Two-colour array capability is based upon stacking materials with different spectral responses on top of each other. The shorter wavelength flux is absorbed in the first layer, which then transmits the longer wavelength flux through the second layer. One such structure—a HgCdTe two-colour device, with two indium bumps per pixel—is illustrated in figure A1.4.46.

Two-colour QWIP detector structures have also been built [27]. For example, figure A1.4.47 shows the excellent imagery in each colour. Note the appearance of the front-held optical filter and the vertically-held hot soldering iron in the two bands. At present, however, imaging systems using two-colour arrays are in limited use. Some considerations have suggested that three-colour FPAs would be more generally useful. Recently, a four-colour QWIP FPA has been demonstrated by stacking different multi-quantum well structures, which are sensitive in 4–5.5, 8.5–10, 10–12, and 13–15.5 μm bands. The 640 × 512 format FPA consists of four 640 × 128 pixel areas which are capable of acquiring images in these bands.

It should be mentioned that hyperspectral arrays are distinguished from multispectral ones in typically having a hundred or more bands. HgCdTe, and other detector materials such as silicon and InSb, have been used in hyperspectral assemblies in the form of two-dimensional arrays with a closely

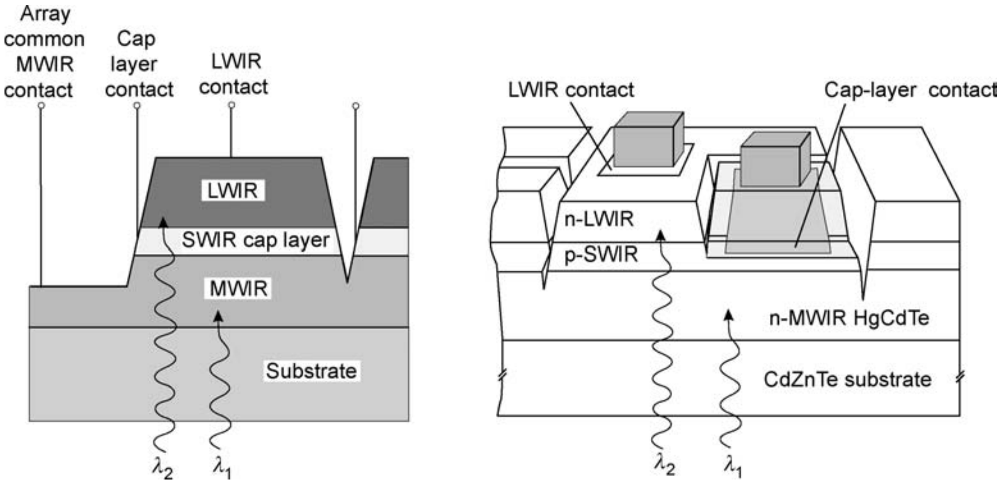

Array common MWIR contact
Cap layer contact
LWIR contact
LWIR
SWIR cap layer
MWIR
Substrate
λ_2
λ_1
LWIR contact
Cap-layer contact
n-LWIR
p-SWIR
n-MWIR HgCdTe
CdZnTe substrate
λ_2
λ_1

Figure A1.4.46. Cross-section of integrated two-colour HgCdTe detectors in an n–p–n layer structure for simultaneous operating mode. A thin p-type barrier separates the two absorbing bands.

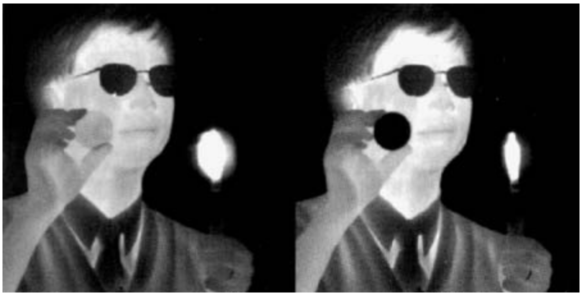


Figure A1.4.47. Simultaneous images from 256 × 256 MWIR/LWIR QWIP FPAs. Note appearance of the front-held filter and the hot soldering iron in the two bands. Reproduced from [28].

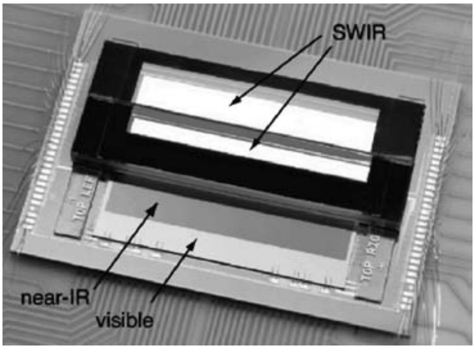

SWIR
near-IR
visible

Figure A1.4.48. Hyperspectral array with 300 bands made from silicon and HgCdTe together with four wedge filters. The array has 100 spectral bands in the range from 0.4 to 1.0 μm, and 200 from 1.0 to 2.5 μm. Reproduced from [18].

packed layout of rows and columns. A prism, grating, or a 'wedged' filter is used to illuminate each row with a different wavelength. Figure A1.4.48 shows an example of a hyperspectral array.

A1.4.5 Conclusions

This chapter provides an overview of the important techniques for detection of optical radiation from the UV, through visible to infrared spectral regions. In the beginning single-point devices are considered, next direct detector systems and advanced techniques including coherent detection, and finally image counterparts containing FPAs are considered. The reader should be able to gain a good understanding of the similarities and contrasts, the strengths and weaknesses of the great number of approaches that have been developed over a century of effort to improve our ability to sense photons. The emphasis is always upon the methods of operation and limitations of different techniques. In addition, currently achieved performance levels are also briefly described.

This chapter offers a rather wide coverage of detection techniques. However, for a full understanding of the technical content, basic courses in electronic devices and circuits, and the very fundamentals of semiconductors and noise is a prerequisite.

References

[1] Donati S 1999 *Photodetectors. Devices, Circuits, and Applications* (New Jersey: Prentice Hall)
[2] Motchenbacher C D and Connelly J A 1995 *Low-noise Electronic System Design* (New York: Wiley)
[3] Vampola J L 1999 Readout electronics for infrared sensors *The Infrared and Electro-Optical Systems Handbook* vol. 3, ed W D Rogatto (Bellingham: SPIE)
[4] Bielecki Z, Kołosowski W, Dufrene R and Borejko M 2003 *Proceedings of 11th European Gallium Arsenide and other Compound Semiconductors Application Symposium*, Munich pp 137–140
[5] Rieke G H 1994 *Detection of Light: From Ultraviolet to the Submillimeter* (Cambridge: Cambridge University Press)
[6] Bielecki Z 2002 Maximisation of signal to noise ratio in infrared radiation receivers *Opto-Electron. Rev.* **10** 209–216
[7] http://opticb.uoregon.edu.
[8] http://www.lasercomponents.de
[9] Skolnik M I 1990 *Radar Handbook* (New York: McGraw-Hill)
[10] Jacobs S F 1988 Optical heterodyne (coherent) detection *Am. J. Phys.* **56** 235–245
[11] Keiser G 1979 *Optical Fiber Communications* (New York: McGraw-Hill)
[12] Barry J R and Lee E A 1990 Performance of coherent optical receivers *Proc. IEEE* **78** 1369–1994
[13] Brun R 1990 Gallium arsenide eyesafe laser rangefinder *Proc. SPIE* **1207** 172–181
[14] Bielecki Z and Rogalski A 2001 *Detection of Optical Radiation* (Warsaw: WNT) (in Polish)
[15] Rogalski A 2000 *Infrared detectors* (Amsterdam: Gordon and Breach)
[16] Rogalski A 2003 Infrared detectors: Status and trends *Prog. Quant. Electron.* **27** 59–210
[17] Rogalski A 2003 Photon detectors *Encyclopedia of Optical Engineering* ed R Driggers (New York: Marcel Dekker) pp 1985–2036
[18] Norton P 2003 Detector focal plane array technology *Encyclopedia of Optical Engineering* ed R Driggers (New York: Marcel Dekker) pp 320–348
[19] http://www.dalsa.com/
[20] Kozlowski L J, Montroy J, Vural K and Kleinhans W E 1998 Ultra-low noise infrared focal plane array status *Proc. SPIE* **3436** 162–171
[21] Dereniak E L and Crowe D G 1984 *Optical Radiation Detectors* (New York: Wiley)
[22] Kozlowski L J, Vural K, Luo J, Tomasini A, Liu T and Kleinhans W E 1999 *Opto-Electron. Rev.* **7** 259–269
[23] Fossum E R 1993 Active pixel sensors: Are CCD's dinosaurs? *Proc. SPIE* **1900** 2–14
[24] Fossum E R and Pain B 1993 Infrared readout electronics for space science sensors: State of the art and future directions *Proc. SPIE* **2020** 262–285
[25] Hewitt M J, Vampola J L, Black S H and Nielsen C J 1994 Infrared readout electronics: a historical perspective *SPIE* **2226** 108–119
[26] Love P J, Ando K J, Bornfreund R E, Corrales E, Mills R E, Cripe J R, Lum N A, Rosbeck J P and Smith M S 2002 Large-format infrared arrays for future space and ground-based astronomy applications *Proc. SPIE* **4486** 373–384
[27] Rogalski A 2003 Quantum well photoconductors in infrared detector technology *J. Appl. Phys.* **93** 4355–4391
[28] Whitaker T 1999 Sanders' QWIPs detect two color at once *Compound Semiconductors* **5** (7) 48–51

A1.5

Propagation along optical fibres and waveguides

John Love

In this Part of the Handbook, an introduction to the description of the transmission of light along dielectric optical fibres and waveguides is presented in terms of a ray analysis for multimode propagation and in terms of a modal analysis for single- and few-mode propagation. Whilst the emphasis will be on fibre propagation, the methods presented here are applicable to any type of waveguide, irrespective of material composition. To cater for a wider range of backgrounds in optics and electromagnetism, there is a low-level introduction leading into more advanced topics. Accordingly, some readers may wish to skip the earlier Sections A1.5.1 and A1.5.2. This Part is relatively self-contained and explanatory, and referencing to other material has been kept to a minimum.

A1.5.1 Historical perspective

A1.5.1.1 Light propagation

Light, as we know, is what we see and comes in a range of colours or, equivalently, wavelengths or frequencies that our eyes detect and our brain interpolates. Whatever we are looking at, whether it is this page or the distant stars, light travels in straight lines from the object to our eyes, regardless of the wavelength. This phenomenon occurs because light is a form of electromagnetic radiation and its propagation when considering waves in a uniform dielectric medium, such as air or vacuum, is described by Maxwell's equations. Sometimes though, light can play tricks on us, such as the apparently sloping water level in the swimming pool shown in figure A1.5.1.

While light travels in straight lines, a finite beam of light, whether it comes from a large-aperture torch or the narrow output face of a micron-size coherent semiconductor laser, tends to spread out with this effect being the more noticeable over short distances with the torch beam and over longer distances with the laser. If we combine this spread with the fundamental need for straight-line propagation, it is easy to see why it is not very convenient to use light in air to transmit information over long distances; it will not go around corners nor even follow the curvature of the earth. On top of this limitation, rain, fog, smoke, buildings, topographical features and other obstacles ensure that the beam will be strongly attenuated because of absorption, reflection and scattering.

These impediments notwithstanding, chains of semaphore repeater stations with large moveable wooden arms on the tops of towers were built by a number of European countries in the eighteenth century on the peaks of hills within direct line of sight of one another, i.e. about 20–25 km apart. These links provided the first purely optical, low-bandwidth, long-distance transmission systems between a country's capital and its ports, primarily for military and administrative applications.

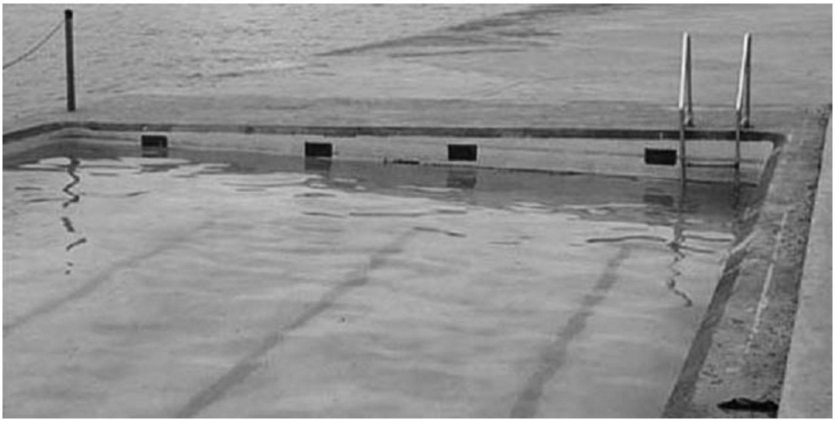


Figure A1.5.1. Swimming pool with an apparently sloping water surface.

A1.5.1.2 Light pipes

However, to provide an optical transmission system with higher bandwidth, a more effective medium was needed that was able to steer light flexibly to its destination and to insulate it from deleterious environmental effects. Following the invention of the laser in the 1950s, early attempts in the 1960s to solve this problem were based on the use of a long, evacuated light pipe, wherein a series of lenses inside the tube periodically refocused the slowly diverging output beam from a laser located at the beginning of the pipe, as shown schematically in figure A1.5.2. The lenses could also be used to steer the beam along new directions to follow the local topography.

Whilst this technology offered a high bandwidth and very low attenuation because light propagation would be predominantly in vacuum, the scheme relied on the maintenance of a good alignment over long distances, regardless of whether the pipe is supported above ground or buried beneath it. This presented a significant challenge because of ground movement due to natural and man-made phenomena. This made it clear that a more flexible and less environmentally sensitive optical guide would be preferable.

A1.5.1.3 Optical fibres

Optical fibres have been known since glass was first discovered several thousands years ago, simply as a consequence of pulling on a piece of heat-softened glass, something that most of us have probably tried in a laboratory at school or university. However, any thoughts on the use of glass fibres for transmitting

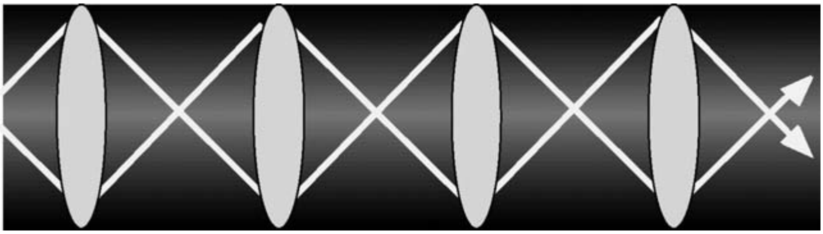


Figure A1.5.2. Schematic of a longitudinal section of a light pipe.

light for information purposes had to wait until the twentieth century and the development of suitable light sources.

The potential for guiding light along a purely dielectric optical waveguide was first demonstrated in public in 1841 by Professor Daniel Colladon at the University of Geneva in Switzerland [1]. Colladon used a jet of water emerging from the side of a barrel and illuminated the jet by focusing sunlight onto it. The water has a higher refractive index of 1.33 compared to the surrounding air with an index of 1.0. Because Snell's laws predict total internal reflection of light from the water-air boundary, the relative indices ensure the sunlight is guided along the water jet. The jet of water curves downwards because of gravity and the strong confining property of the jet determines that much of the light follows its curved path. Using more modern light sources, such as a coloured diode, enhances the appearance of the effect. We now know that Colladon's experiment is equivalent to light propagation along a highly multimode, bent optical waveguide with water as the core and air as the cladding.

Glass fibres were developed spasmodically over the next 100 years more as a curiosity than a technology solution. The first serious application of glass fibres to light transmission was motivated through medicine leading to the development of the endoscope in the 1950s with achievable transmission over one or two metres through a dense bundle of very thin fibres. However, there were two existing limitations that inhibited the use of the endoscope fibres for long-distance communications. Firstly, early fibres were single material and because light is distributed over the entire cross-section of the material, it was susceptible to transmission loss wherever the single material touched external supports. Secondly, the glass was very impure and therefore strongly absorbed and scattered light over the longer distances. One can get an idea of the degree of light absorption by looking through a sheet of window glass sideways, as illustrated in figure A1.5.3. Even a very low level of impurities in the glass, of the order of parts per million, results in a very high extinction level for the propagating light after only a few metres.

The transition from Colladon's water jet to a modern long-distance, low-loss optical communications system relied on (a) the development of two-layer, high-low-index glass optical fibres that not only transmit light but also confine it to within the central glass core well away from the outer surface and, more importantly, (b) the development of techniques for fabricating fibres with extremely low light loss from very pure materials.

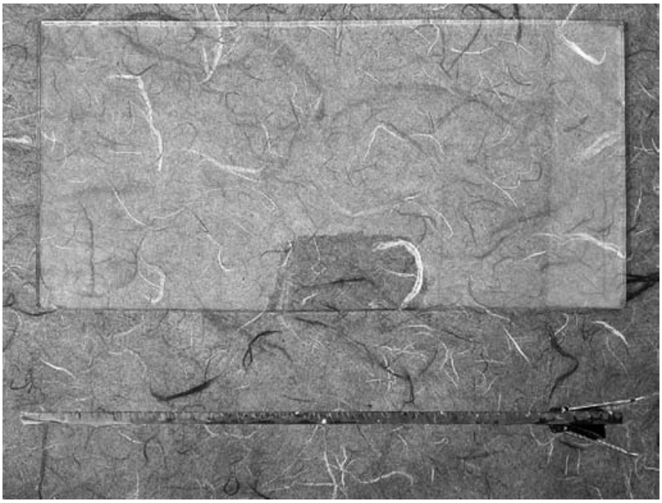


Figure A1.5.3. Top and side views of a sheet of ordinary window glass.

The first limitation was overcome by using the now-standard core-plus-cladding fibre to locate the propagating light around the fibre axis and isolate it from the fibre coating and other external influences. The second problem required the development of new fabrication techniques to produce the very pure silica-based glasses required for low-loss fibres. These goals were enshrined by the classic paper of Kao and Hockham [2], published in 1966, and generally regarded as the catalyst for the ensuing optical telecommunications revolution.

The first such fibres that appeared in the 1960s were multimode and it was not until the 1970s that the goal of low-loss, single-mode fibres and cables with much higher bandwidths than multimode fibres was realized. Contemporary fibre cables for terrestrial and submarine applications are shown in figure A1.5.4.

A1.5.1.4 Scope

In this Part of the Handbook we develop the analysis of propagation along optical fibres and other waveguiding structures to produce a basic description of light transmission. The practical development of fibres was paralleled by major theoretical developments. In the 1960s and early 1970s, much effort was extended to producing increasingly sophisticated ray-tracing and local plane-wave techniques that could adequately describe and quantify propagation and loss mechanisms in multimode fibres. Ray tracing is an accurate technique for multimode fibres because the relatively large core ensures that diffraction effects associated with the relatively short but finite wavelengths used in communications are extremely small compared to the light-guiding effect based on the variation in the transverse refractive-index profile. Conversely, the small core size associated with single-mode fibres requires a full electromagnetic analysis of propagation because diffraction effects become comparable with the confining effect of the profile (see section 10.2 of Ref. [3]).

The early ray-tracing and electromagnetic analyses of fibres concentrated predominantly on analytical solutions of the governing ray-tracing and Maxwell equations because of the limited capability of early computers. Initially, these solutions necessarily relied on refractive-index profiles, together with fibre and waveguide cross-sectional geometries, for which analytical solutions of these equations were available in terms of simple or special mathematical functions. Commercial software routines were available for the quantification of some of these special functions, otherwise home-based software had to be developed.

Over the last 20 years, this situation has been almost completely reversed with the ready availability of commercially developed software, particularly that for determining ray tracing and electromagnetic propagation for a wide range of fibre and waveguide profiles and geometries. It is fair to say that this

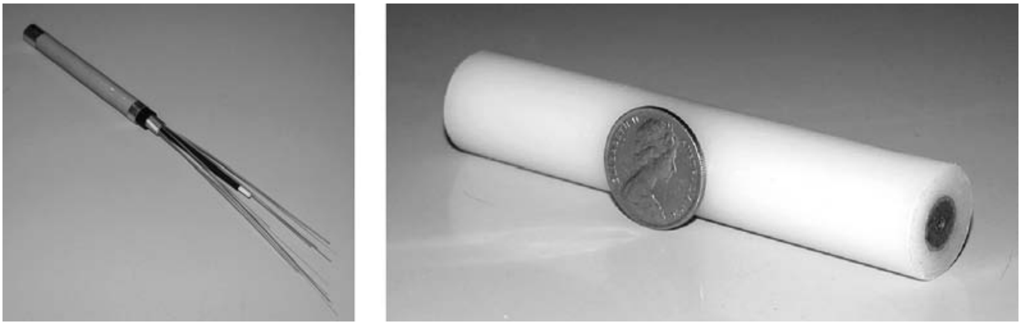


Figure A1.5.4. Terrestrial (left) and deep-ocean submarine (right) single-mode fibre cables used in long-distance optical communications systems.

evolution has considerably simplified the development and understanding of specialized fibre and waveguide designs, such as holey fibres and photonic bandgap waveguides, by moving the emphasis to the exploitation of physical phenomena for light guidance rather than being limited by the shortcomings of analytical and numerical techniques.

A1.5.2 Light propagation, plane waves and rays

Before launching into the different descriptions of light propagation and guidance along fibres and waveguides, it is helpful and insightful to examine the propagation of electromagnetic plane waves in free space as a precursor to their development into rays and modes. Light rays, whether propagating in straight lines or along curved trajectories, are formally the solution of Maxwell's equations in the limit of zero wavelength, but for practical purposes can be thought of as local plane waves propagating with a small but finite wavelength. This section provides an elementary background to the more advanced material presented in Sections A1.5.3 and A1.5.5.

A1.5.2.1 Plane waves

Consider an infinite, unbounded medium of uniform, real refractive index n that is loss-less, i.e. non-absorbing and non-scattering. Introduce the triad of Cartesian axes O-xyz, shown in figure A1.5.5, such that the z-axis defines the direction of propagation.

Assume an artificial, infinitely extended monochromatic (single-frequency) uniform light source in the *x–y plane* with angular frequency ω or, equivalently, wavelength $\lambda = 2\pi c/\omega$, where c is the speed of light in vacuum. An electromagnetic plane wave can propagate parallel to the z-axis with vector electric $\boldsymbol{E}$ and magnetic $\boldsymbol{H}$ fields that are everywhere uniform and have only a sinusoidal dependence on distance z and time t. Accordingly a simple solution of Maxwell's equations predict that these fields have the forms:

$$\boldsymbol{E}(x, y, z, t) = \boldsymbol{e}\exp\{\mathrm{i}(knz - \omega t)\}; \quad \boldsymbol{H}(x, y, z, t) = \boldsymbol{h}\exp\{\mathrm{i}(knz - \omega t)\} \tag{A1.5.1}$$

where $k = 2\pi/\lambda$ is the wavenumber, kn is the propagation constant, and $\boldsymbol{e}$ and $\boldsymbol{h}$ are *constant* orthogonal vectors. Note that the propagation constant is a continuous function of the source wavelength and decreases as the wavelength increases.

These fields propagate parallel to the z-direction with a *phase velocity* $v_{\mathrm{ph}} = \omega/kn = c/n$, i.e. the speed of light in a medium of refractive index n, which is a constant *independent* of the source wavelength or frequency, provided the medium itself is not dispersive. The forward direction of propagation corresponds to the negative sign in the exponent of equation (A1.5.1). Furthermore, the power in the

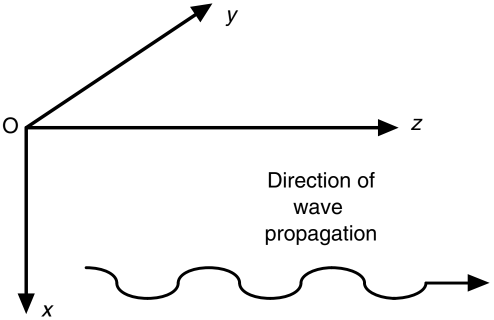


Figure A1.5.5. Orientation of Cartesian O-xyz axes.

plane wave propagates at the *group velocity*, $v_g = d\omega/dk = \omega/kn = c/n$, i.e. identical to the group velocity for a plane wave.

A plane wave is an unphysical entity because it has constant amplitude everywhere and therefore the total power propagating in the mode is infinite. Nevertheless, it provides the simplest example of light propagation in a uniform medium. It will be seen in Section A1.5.3 that the introduction of a core into the uniform medium model modifies the nature of plane-wave propagation in a constructive manner.

A1.5.2.2 Polarization

The two constant vectors $\boldsymbol{e}$ and $\boldsymbol{h}$ each have only one *non-zero* Cartesian field component transverse to the z-direction of propagation. These components lie in the x–y plane.

Relative to the Cartesian axes, these vectors can be chosen to have either of the following two orthogonal forms:

$$\boldsymbol{e} = (e, 0, 0) \quad \boldsymbol{h} = (0, h, 0) \quad \text{or} \quad \boldsymbol{e} = (0, e, 0) \quad \boldsymbol{h} = (-h, 0, 0) \tag{A1.5.2}$$

where e and h are scalar constants. Note that the second choice of components is simply the first choice rotated 90 degrees about the z-axis. Either of these two solutions consists of electric and magnetic fields that are orthogonal to the z-direction of propagation and are also orthogonal to one another. Accordingly they form *orthogonal triads* as shown in figure A1.5.6.

The direction of the electric field vector $\boldsymbol{E}$ in each case defines the *polarization* of the plane wave, i.e. it is either x-polarized or y-polarized, respectively. The description of the plane wave for either polarization can be thought of as a triad of vectors ($\boldsymbol{E}$, $\boldsymbol{H}$, and the z-axis or direction of propagation) propagating parallel to the z-axis at a speed equal to the phase velocity v_{ph}.

A1.5.2.3 Local plane waves, rays and waves

Plane waves are associated with an infinite medium of uniform index n, and they travel in straight lines and everywhere have the same electromagnetic description. However, if the refractive index of the medium varies with position so that it is *graded*, this description is modified. The plane wave can be replaced by the concept of the *local plane wave*, which describes propagation in a small region of space in terms of a wave whose fields, phase and group velocities, and polarization are determined by the local value and variation of the index.

When propagation is described in terms of ray tracing for the multimode waveguides and fibres in Section A1.5.5, there is a simple relationship between rays and waves. A ray represents the local direction of propagation of a plane wave in a uniform medium or a local plane wave in a graded

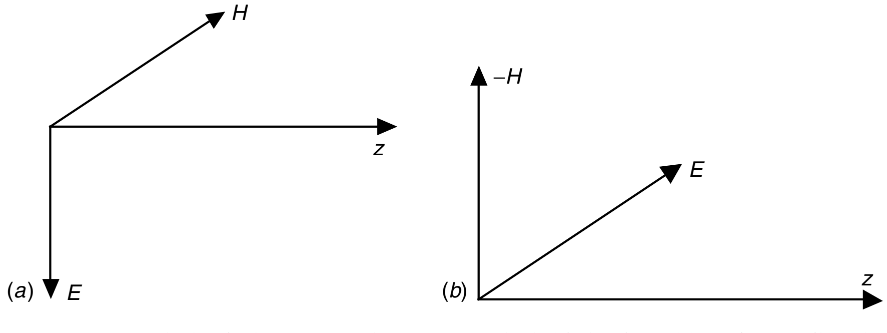


Figure A1.5.6. Polarized plane wave electromagnetic fields in the x (a) and y (b) directions.

medium. Further, the ray direction also determines the direction of local power flow. Although this propagation description is based on the zero-wavelength limit of Maxwell's equations, it provides an accurate description for the relatively short wavelengths encountered in optical communications.

A1.5.3 Electromagnetic propagation in non-uniform dielectric media

A1.5.3.1 Maxwell's equations and monochromatic sources

Propagation in dielectric media is governed by Maxwell's equations (a portrait of Maxwell is given in figure A1.5.7) so that in the absence of currents and charges the equations take the four-dimensional spatial-temporal forms:

$$\nabla \times \boldsymbol{E} = \frac{\partial \boldsymbol{B}}{\partial t}; \quad \nabla \times \boldsymbol{H} = -\frac{\partial D}{\partial t} \tag{A1.5.3}$$

where $\boldsymbol{E} = \boldsymbol{E}(x, y, z, t)$ is the vector electric field and $\boldsymbol{H} = \boldsymbol{H}(x, y, z, t)$ is the vector magnetic field. The temporal dependence is denoted by t and the spatial dependence can be expressed in terms of any convenient orthogonal coordinate system, e.g. Cartesian coordinates (x, y, z) for slab waveguides, cylindrical polar coordinates (r, ϕ, z) for circular fibres, etc. The magnetic induction vector $\boldsymbol{B}$ and the displacement vector $\boldsymbol{D}$ are related to the magnetic and electric fields, respectively, by

$$\boldsymbol{B} = \mu \boldsymbol{H}; \quad \boldsymbol{D} = \varepsilon \boldsymbol{E} = \varepsilon_0 n^2 E \tag{A1.5.4}$$

where μ is the magnetic permeability, $\varepsilon = \varepsilon(x, y, z)$ is the dielectric constant, and $n = n(x, y, z)$ is the refractive index distribution. For optical materials, μ normally takes its free-space value μ_0, and ε_0 is the free-space dielectric constant. For fibres and waveguides, the refractive index is normally assumed to be uniform (z-independent) along the fibre or waveguide and is also taken to be independent of the field amplitude, i.e. ignoring non-linear material effects as well as the source wavelength when ignoring

Figure A1.5.7. James Clerk Maxwell, 1831–1879.

material dispersion. Material dispersion is addressed in Section A1.5.6.4 and non-uniform, z-dependent propagation is discussed in Section A1.5.9.3.

Monochromatic sources

In modelling the excitation from lasers or diodes when used as sources for fibres and waveguides, it is initially assumed that their output is exactly sinusoidal and monochromatic with a fixed wavelength. Practical sources, however, have an output with a finite but small *spectral width*. For lasers this width is typically of the order of a nanometres or less. The effect of this width is of paramount importance when considering pulse dispersion (see Section A1.5.6).

Accordingly, we ascribe an angular frequency ω to the monochromatic source such that each component of the electromagnetic field is assumed to contain the implicit sinusoidal dependence $\exp(-\mathrm{i}\omega t)$. The choice of sign is arbitrary, but here is taken to be negative for convenience. If λ denotes the corresponding free-space wavelength and k is the free-space wavenumber, then these quantities are related by the expressions

$$k = \frac{2\pi}{\lambda}; \quad \lambda = \frac{2\pi c}{\omega}; \quad c = \frac{\omega}{k} \tag{A1.5.5}$$

where c is the free-space speed of light ($3 \times 10^8\,\mathrm{ms}^{-1}$). Using equations (A1.5.3–A1.5.5), the monochromatic time dependence enables us to recast Maxwell's equations so that the spatial dependence is governed by

$$\nabla \times \boldsymbol{E} = \mathrm{i}\omega\mu_0 \boldsymbol{H} = \mathrm{i}\left(\frac{\mu_0}{\varepsilon_0}\right) kH; \quad \nabla \times \boldsymbol{H} = \mathrm{i}\omega\varepsilon_0 n^2 \boldsymbol{E} = -\mathrm{i}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} kn^2 \boldsymbol{E} \tag{A1.5.6}$$

where $\boldsymbol{E} = \boldsymbol{E}(x, y, z)$ and $\boldsymbol{H} = \boldsymbol{H}(x, y, z)$ and implicitly contain the time dependence $\exp(-\mathrm{i}\omega t)$.

A1.5.3.2 Translational invariance, longitudinal and transverse fields

For modelling fibres and waveguides, it is usual to assume that (a) the refractive index profile and the cross-sectional geometry do not vary with longitudinal distance z and (b) the fibre or waveguide is straight and essentially infinitely long. In this situation the fibres and waveguides have *translational invariance* and hence the z-dependence in Maxwell's equations becomes separable from the transverse dependence. Accordingly we may set

$$\boldsymbol{E}(x, y, z) = \boldsymbol{e}(x, y)\mathrm{e}^{\mathrm{i}\beta z}; \quad \boldsymbol{H}(x, y, z) = \boldsymbol{h}(x, y)\mathrm{e}^{\mathrm{i}\beta z} \tag{A1.5.7}$$

where $\boldsymbol{e}(x, y)$ and $\boldsymbol{h}(x, y)$ are vector expressions that denote the transverse field dependence and the parameter β in the exponential or phase term is called the *propagation constant*. The choice of Cartesian (x, y) coordinates is appropriate for analysing slab waveguides, while polar coordinates (r, ϕ) are more appropriate for fibres.

It is sometimes convenient for analytical and other purposes to split the fields into longitudinal and transverse components. This is equivalent to decomposing the vector dependences of $\boldsymbol{e}$ and $\boldsymbol{h}$ into transverse (subscript 't') and longitudinal (subscript 'z') components, respectively perpendicular and parallel to the z-axis, according to

$$\boldsymbol{e}(x, y) = \boldsymbol{e}_{\mathrm{t}}(x, y) + e_{\mathrm{z}}(x, y)\hat{\boldsymbol{z}}; \quad \boldsymbol{h}(x, y) = \boldsymbol{h}_{\mathrm{t}}(x, y) + h_{\mathrm{z}}(x, y)\hat{\boldsymbol{z}} \tag{A1.5.8}$$

where $\boldsymbol{e}_{\mathrm{t}}$ and $\boldsymbol{h}_{\mathrm{t}}$ are vector quantities, e_{z} and h_{z} are scalar quantities, and $\boldsymbol{z}$ is the unit vector parallel to the fibre axis.

A1.5.3.3 Power density and flow

For a monochromatic source with a sinusoidal time variation, the power flow density and direction at any position in space is determined by the time-averaged Poynting vector $\boldsymbol{S}$

$$\boldsymbol{S} = \frac{1}{2}\mathrm{Re}(\boldsymbol{E} \times \boldsymbol{H}^*) \qquad \text{(A1.5.9)}$$

where * denotes complex conjugate, $\boldsymbol{x}$, the vector cross product and Re is the real part. Loss-less propagation of light in fibres and waveguides is equivalent to the direction of $\boldsymbol{S}$ anywhere in the cross-section being parallel to the z-axis, i.e. $\boldsymbol{S}$ had only a z-component S_z. This property can be readily verified for the specific examples considered in Section A1.5.4. Finally, the total guided power flow is determined by integrating S_z over the infinite cross-section of the fibre or waveguide.

A1.5.3.4 Boundary conditions

For dielectric fibres and waveguides in the absence of currents, the boundary conditions for any bound solution of Maxwell's equations between regions of different index can be stated as follows:

- continuity of all three components of the magnetic field at any interface;
- continuity of the two *tangential* components of the electric field at any interface;
- continuity of the *normal* component of the displacement vector at any interface;
- the electric and magnetic fields decrease exponentially to zero at infinite distance from the fibre or waveguide axis.

In the case of an interface between one region of uniform index and a second region of varying index, where the index values are equal on the interface, all six components of the electric and magnetic fields will be continuous across the interface.

A1.5.3.5 Electromagnetic normal modes

Starting with an unbounded uniform index medium, consider an infinitely long uniform slab or cylinder of material of higher but uniform *core* refractive index n_{co} that is introduced into the infinite medium parallel to the z-axis. If the surrounding infinite medium is now referred to as the *cladding* with index $n = n_{\mathrm{cl}} < n_{\mathrm{co}}$, then the core and cladding constitute a dielectric slab waveguide or fibre with the step refractive index profile cross-section shown in figure A1.5.8. Although practical fibres or waveguides

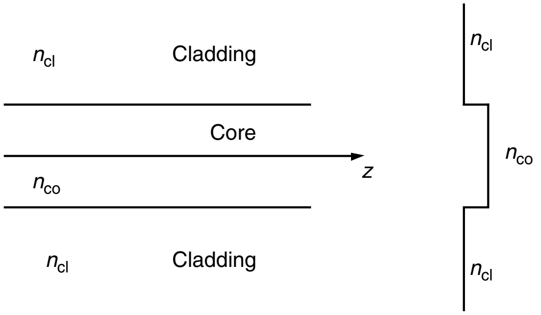


Figure A1.5.8. Step refractive index profile for a fibre or waveguide.

normally have more complex geometrical and profile structures in the cross-section, the step profile is the simplest profile to analyse that is of some practical interest and is used in single-mode telecommunications fibres such as SMF28.

Propagation along a fibre or waveguide is governed formally by Maxwell's equations, but some of the salient physical features can be elucidated by considering propagation of electromagnetic waves along this structure parallel to the z-axis.

The simple plane-wave description of propagation in the uniform medium discussed in Section A1.5.2 provides a conceptual base, but is now modified in a number of ways because of the presence of the waveguide structure.

(a) The uniformity of the structure in the z-direction or *translational invariance* ensures that waves propagate with a periodic longitudinal and temporal dependence and therefore a well-defined phase velocity ν_{ph};

(b) The value of the propagation constant is dependent on both the source wavelength λ and the length parameters of the core geometry and the refractive index values;

(c) The propagation constant is no longer uniquely specified and can take *one or more discrete values* for a given set of values of all the above parameters;

(d) The transverse dependence of the electromagnetic field vectors $\boldsymbol{e}$ and $\boldsymbol{h}$ for each discrete value of the propagation constant is now *spatially dependent* and varies with position (x, y) in the cross-section;

(e) The transverse electric and magnetic fields are no longer uniform, but are concentrated within and close to the core and decrease exponentially to zero at infinite distance from the axis;

(f) The *infinite* power travelling in the infinite cross-section of the plane wave is replaced by the *finite* power propagating along the waveguide in each propagation state;

(g) Each discrete propagation state is known as a *bound mode* or *normal mode* and is a solution (or eigenfunction) of the *electromagnetic boundary value* problem for Maxwell's equations for the waveguide or fibre.

Mechanical analogue

Normal modes are two-dimensional electromagnetic vibrations in the cross-section of the waveguide or fibre and are analogous to the mechanical vibrations of a flexible membrane fixed along its periphery, e.g. a drum. Given the source frequency, each electromagnetic vibration state corresponds to a discrete value of the propagation constant β. Each value of β corresponds to a normal or bound mode where its total longitudinal power flow (parallel to the z-axis) is constant and where its electromagnetic fields decrease exponentially with increasing distance from the z-axis and vanish at infinity.

The values of β are the discrete solutions of an *eigenvalue equation*, derived from solutions of Maxwell's equations together with appropriate boundary conditions. If the fibre core radius ρ or the waveguide cross-section 2ρ is small enough, typically a few microns for a silica-based fibre or waveguide, then as will be shown in Sections A1.5.4.4 and A1.5.4.5 there is only one solution for β, i.e. a *single-mode* waveguide or fibre.

A1.5.3.6 Mode orthogonality, normalization and orthonormal modes

Each bound mode is orthogonal to all other bound modes on a uniform, straight waveguide. This means that if only one mode is excited, it cannot excite any other modes as it propagates. Mathematically, *orthogonality* between the *j*th and *k*th bound modes is expressed by the vanishing of the integral of a triple scalar product of their vector fields over the infinite cross-section of the fibre or waveguide

$$\int_{A_\infty} \boldsymbol{e}_j \times \boldsymbol{h}_k^* \cdot \hat{z}\, \mathrm{d}A = 0 \tag{A1.5.10}$$

where A_∞ is the infinite cross-section, $\hat{z}$ is the unit vector parallel to the axis, and $*$ denotes the complex conjugate.

If we assume that the refractive index profile $n(x,y)$ is independent of field intensity, as is the case for *linear materials*, then Maxwell's equations constitute, in general, a set of coupled linear equations. The solutions of these equations together with the boundary conditions can only determine the fields for each mode to within an *arbitrary amplitude constant*. The value of this constant for each mode is normally determined from the source of excitation as discussed in Section A1.5.8.3.

For some special applications, such as in tapers, it is useful to be able to specify the value of this constant in an unambiguous manner independent of the source of excitation. This can be achieved by using the *normalization N* of a mode. This scalar quantity

$$N = \int_{A_\infty} \boldsymbol{e} \times \boldsymbol{h}^* \cdot \hat{z}\, \mathrm{d}A \tag{A1.5.11}$$

is defined in terms of the integral over the infinite cross-section of the triple scalar product of the electric field $\boldsymbol{e}$, magnetic field $\boldsymbol{h}$, and the unit vector parallel to the z-axis.

An *orthonormal mode* is then defined to be a normal mode with electric and magnetic field amplitudes such that it has unit normalization, i.e. $N = 1$. If the fields $\boldsymbol{e}$ and $\boldsymbol{h}$ of the bound mode are replaced by the following quantities

$$\hat{\boldsymbol{e}} = \frac{\boldsymbol{e}}{N^{1/2}}; \quad \hat{\boldsymbol{h}} = \frac{h}{N^{1/2}} \tag{A1.5.12}$$

then substitution into equation (A1.5.11) gives $N = 1$ and the fields are orthonormal.

A1.5.3.7 Modal phase and group velocities

The longitudinal and temporal dependence of the fields of a mode is contained in the common phase factor $\exp[\mathrm{i}(\omega t - \beta z)]$. A constant value of this expression defines a *phase front*, which is a plane of constant phase orthogonal to the z-axis. For a constant phase value, z varies linearly with t, the constant of proportionality being the ratio $\omega/\beta = v_{\mathrm{ph}}$ or the *phase velocity* of the mode. This is the speed at which the phase front propagates.

For pulse propagation, the speed with which energy is transmitted by a mode is given by the *group velocity* with the standard definition $v_{\mathrm{gr}} = \mathrm{d}\omega/\mathrm{d}\beta$. Since different modes have different values of the propagation constant β for the same source frequency, it follows that the phase and group velocities will also differ between modes. The modal group velocity determines pulse dispersion, as discussed in Section A1.5.6.

A1.5.3.8 Propagation constant, effective index and cut-off

For a given refractive index profile, the propagation constant β for every bound mode occupies a defined range of values in terms of the maximum n_{co} and minimum n_{cl} refractive index values and the source wavelength. If we assume that the modal phase velocity v_{ph} must lie between the maximum and minimum values of the speed of light in the cladding and core, i.e. $c/n_{co} < v_{ph} < c/n_{cl}$, then on setting $c = \omega/k$ it follows that

$$kn_{cl} < \beta < kn_{co} \tag{A1.5.13}$$

where $k = 2\pi/\lambda$ is the free-space wavenumber. This result applies to any waveguide independent of the number of bound modes that can propagate.

Sometimes it is more convenient physically to discuss propagation constant values in terms of an *effective index* value. If the effective index n_{eff} is defined through the relationship $\beta = kn_{eff}$, then equation (A1.5.13) is replaced by

$$n_{cl} < n_{eff} < n_{co} \tag{A1.5.14}$$

In other words, each bound mode has an effective index value that must lie between the minimum and maximum refractive index values.

When $\beta = kn_{cl}$ or, equivalently, $n_{eff} = n_{cl}$, a mode becomes *cut off* and for $\beta < kn_{cl}$ or $n_{eff} < n_{cl}$, a mode is said to be *below cut-off*.

A1.5.3.9 Waveguide, fibre and modal parameters

Waveguide and fibres are commonly characterized in terms of a number of standard dimensionless parameters that combine various basic parameters. In the following definitions, it is assumed that n_{co} denotes the uniform core index in the case of a step profile or the maximum core index in the case of a graded profile, and n_{cl} denotes the uniform cladding index. The source wavelength is λ, $k = 2\pi/\lambda$ is the free-space wavenumber and ρ is the core radius in the case of a circular fibre or the core half-width of a slab or square-core waveguide. Note that some other definitions assume that ρ is the full width of a slab or square-core waveguide.

Relative index difference

The uniform cladding index n_{cl} and the uniform or maximum core index n_{co} are often combined to define the *relative index difference* Δ

$$\Delta = \frac{n_{co}^2 - n_{cl}^2}{2n_{co}^2} \cong \frac{n_{co} - n_{cl}}{n_{co}} \tag{A1.5.15}$$

where the second expression is obtained by factorizing the numerator in the first expression and assuming that $n_{co} \sim n_{cl}$. This representation is appropriate within the weak guidance approximation discussed in Section A1.5.4.

Numerical aperture

A measure of the light-capturing capacity of a waveguide or fibre is provided by the *numerical aperture* or NA, which is defined by

$$\mathrm{NA} = (n_{co}^2 - n_{cl}^2)^{1/2} = n_{co}2\Delta^{1/2} \tag{A1.5.16}$$

A physical interpretation of numerical aperture is presented in Section A1.5.5.6.

Waveguide and fibre parameter or frequency

A second parameter V combines all the key parameters of a waveguide or fibre into a single normalized quantity known as the *waveguide* or *fibre parameter* or *frequency* that can be expressed in a number of equivalent ways

$$V = \frac{2\pi\rho}{\lambda}(n_{\mathrm{co}}^2 - n_{\mathrm{cl}}^2)^{1/2} = k\rho(n_{\mathrm{co}}^2 - n_{\mathrm{cl}}^2)^{1/2} = k\rho n_{\mathrm{co}}(2\Delta)^{1/2} \tag{A1.5.17}$$

This is an important parameter for its value determines, in particular, whether a fibre or waveguide is single mode, as discussed in Sections A1.5.4.4 and A1.5.4.5.

Modal parameters

In the examples of modal analyses of waveguides and fibres presented in Sections A1.5.4.4 and A1.5.4.5, respectively, it is convenient to introduce normalized *modal parameters* U and W for the core and cladding of a fibre or waveguide, respectively, which incorporate the propagation constant or equivalently the effective index according to

$$U = k\rho(n_{\mathrm{co}}^2 - \beta^2)^{1/2} = k\rho(n_{\mathrm{co}}^2 - n_{\mathrm{eff}}^2)^{1/2}; \quad W = k\rho(\beta^2 - n_{\mathrm{cl}}^2)^{1/2} = k\rho(n_{\mathrm{eff}}^2 - n_{\mathrm{cl}}^2)^{1/2} \tag{A1.5.18}$$

It then follows from the definition of V in equation (A1.5.17) that

$$U^2 + W^2 = V^2 \tag{A1.5.19}$$

The values of U and W are necessarily discrete for bound modes.

A1.5.3.10 Radiation modes, leaky modes and super-modes

The range of propagation constant values for bound modes satisfies equation (A1.5.14). If $\beta > kn_{\mathrm{co}}$, propagation is not possible, but if $\beta < kn_{\mathrm{cl}}$, it is possible to analyse propagation of the unguided field in the fibre or waveguide in using one of three different descriptions. These descriptions depend on the particular physical model employed and whether the cladding is unbounded or finite, but large, compared to the core size. In each case, the particular method is focused on determining the propagation characteristics of light within the waveguide that is not guided by the bound modes.

Radiation modes

In the case of a fibre or waveguide with an *unbounded cladding*, there are no bound modes with propagation constants in the range $0 < \beta < kn_{\mathrm{cl}}$ but, instead, a continuum of *radiation modes* can be derived, each mode having a continuously varying propagation constant value in this range. These modes can be used to analyse non-guided propagation in terms of an integration over each radiation mode and a sum over the integrals for different radiation modes (see chapter 25 of Ref. [4]). This approach is analytically complex and in view of the ready availability of vector-based beam propagation methods (BPM) and other techniques, a numerical analysis may be easier to implement.

Leaky modes

The continuum of radiation mode solutions can be approximated by a discrete summation of so-called *leaky modes*. A leaky mode is defined by the analytic continuation of a bound mode to propagation constant values beyond the mode's cut-off, i.e. for $\beta < kn_{\mathrm{cl}}$. However, unlike a bound mode, where the power flow is everywhere parallel to the z-axis, the local power flow is at an angle to the axis so that power flows away from the core as the mode propagates. This divergence corresponds to a complex value of the leaky mode propagation constant. Furthermore the power of a leaky mode is unbounded, so that a proper quantitative analysis of propagation cannot be undertaken, but nevertheless leaky modes can provide useful physical insight into propagation characteristics when radiation is present (see chapter 24 of Ref. [4]).

Super-modes

The cladding of any practical waveguide is necessarily finite in cross-section and may be surrounded by air or by a protective coating in the case of a fibre. In either case, it is normally possible to describe the transient field of a waveguide excited by a source using a superposition of the complete set of bound modes of the complete core-cladding-air structure. To distinguish these modes from the modes guided by the core-cladding refractive index profile, the former are often referred to as *super-modes*. The dimension of the complete cross-section is relatively large compared to that of the core ensuring that the number of super-modes is very large. In the limit of an infinitely thick cladding, the super-mode and radiation mode solutions approach one another [5].

Holey fibres

Currently there is significant interest in the light-guiding properties of *holey fibres*, which have certain attributes that are quite different to those of solid material fibres. Holey fibres differ from conventional fibres in that they are normally fabricated from a single material, such as silica or a polymer with a uniform refractive index, compared with a conventional fibre that has an index contrast between the core and the cladding materials.

Light guidance along holey fibres depends on the presence of concentric circular arrays of small longitudinal holes about the fibre axis (figure A1.5.9). Each ring of holes can be regarded as defining an annular region in which the average index is smaller than the material index and hence provides an effective index contrast with the material region around the fibre core and outside the ring.

However the core and ring of holes do not support any bound modes. The concentric region outside of the ring of holes is of the same index as the core and this enables any modal field within the holes to gradually leak from the centre across the ring of holes to the outer region. In other words, holey fibres only support *leaky modes*. Nevertheless, by judicious choice of the size, number and distribution of air holes, it is possible to design a fibre whereby the fundamental mode has virtually zero attenuation leakage but all higher-order modes have relatively large leakage rates so that their fields rapidly disappear from the central region as they propagate along the fibre. For practical purposes the holey fibre then behaves like a single-mode fibre, even over very long distances.

Photonic band-gap or *crystal* fibres also guide light, but the guidance mechanism is different to that of holey fibres. The physical basis for guidance along such fibres relies on a special arrangement of many rings of holes about the z-axis such that the rings provide a band gap in the radial direction, i.e. a barrier to the propagation of light away from the axis. One way to think of the band-gap effect qualitatively is to smear out the index contrast between the holes and the fibre material in each ring into an average reduced index. Then the quasi-periodic radial variation between the rings and the fibre material constitutes an effective radial Bragg reflection grating. At the effective Bragg wavelength in the fibre cross-section, the grating inhibits propagation and sets up a radial evanescent field that decreases radially outward. In other words, for suitably chosen fibre parameters the layers of rings can

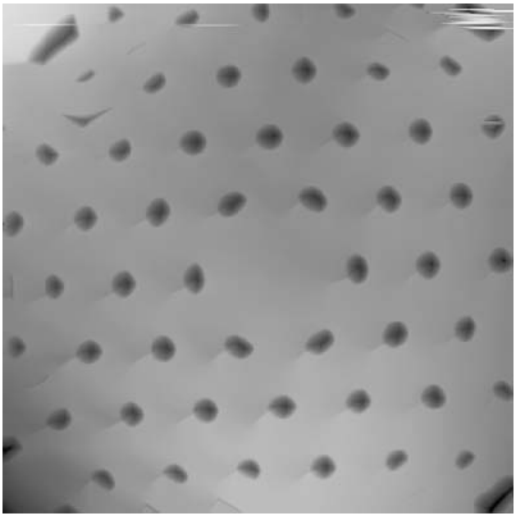


Figure A1.5.9. Cross-section of a holey fibre showing the solid core and surrounding rings of air holes.

support the evanescent field of the fundamental mode propagating along the fibre with the majority of its field close to the fibre axis. Chapter B10 provides a detailed description and analysis of both kinds of fibre.

A1.5.3.11 Polarization, mode nomenclature and birefringence

When a mode propagates along a fibre or waveguide, the magnitude and direction of the transverse components of its electric and magnetic fields remain fixed and thereby define the orientation of the modal field. The *polarization* of a mode is defined by the direction of the transverse electric field vector at each position in the waveguide or fibre cross-section. Generally this direction will vary with position so that contours of the field direction are normally curved, but in the case of two-dimensional slab waveguides, *all modes* have a transverse electric field that is parallel to a fixed direction, i.e. each mode is *plane polarized*. A similar situation pertains to the modes of weakly guiding fibres where transverse electric fields are also plane polarized.

Mode nomenclature

For two-dimensional symmetric slab waveguides, there are two distinct classes of modes, depending on their polarization. Transverse electric or TE_j modes have only a single transverse electric field component and transverse magnetic or TM_j modes have only a single transverse magnetic field component, where $j = 0, 1, 2, 3 \ldots$ denotes the mode order. The even and odd values of j denote modes with even and odd field symmetry, respectively. As the value of j increases, the number of extrema in the field patterns also increases.

In the case of circular fibres the situation is more complex because of the radial and azimuthal directions. There is a class of cylindrically symmetric TE_j and TM_j modes that have electric fields with a single radial or azimuthal transverse component, respectively, where $j = 1, 2, 3 \ldots$. In addition, there is the class of hybrid HE_{ij} and EH_{ij} modes that can be regarded as linear combinations of both TE and TM field components and therefore have a more complex electric field structure in the fibre cross-section.

Here the subscript 'i' relates to the order of the azimuthal angular dependence and 'j' refers to the radial order, the value of which increases with the number of extrema in the field. The azimuthal dependence of each mode has a $\cos(m\varphi)$ or $\sin(m\varphi)$ variation where $m = 0, 1, 2, \ldots$.

In the case of weakly guiding circular fibres, a second complimentary mode nomenclature is commonly used whereby each mode is labelled as LP_{ij}, the 'LP' denoting *linearly polarized*. The subscripts 'i' and 'j' denote azimuthal and radial orders, respectively, and in terms of HE modes in the weak guidance approximation the relationship can be expressed as $\mathrm{LP}_{ij} \leftrightarrow \mathrm{HE}_{i+1,j}$.

Within the two nomenclatures, the fundamental mode is equivalent to the HE_{11} mode for arbitrary index difference circular fibres and to either the HE_{11} or LP_{01} mode for weakly guiding fibres. For other waveguide and fibre geometries there is no generally accepted nomenclature for categorizing modes.

Birefringence

On a circular weakly guiding fibre, the fields of the fundamental mode are rotationally invariant about the fibre axis, so that any pair of orthogonal directions can be chosen for its two polarization states that have identical propagation constants. In this situation the fundamental mode is said to be *degenerate*.

However, in the case of non-circular fibres, such as the elliptical core fibre shown in figure A1.5.10, this is no longer the situation. Working within the weak guidance approximation, the two polarization directions of the planar transverse electric field are parallel to one of the *optical axes* of the fibre. For the elliptical cross-section the optical axes coincide with the major x-axis and minor y-axis and the fundamental mode has respective propagation constants β_x and β_y. Since $\beta_x > \beta_y$ the fundamental mode is *non-degenerate* and the fibre is said to be *birefringent* because of the difference in propagation constant for the two polarization states introduced by the non-circular core geometry.

Birefringence is a measure of the difference in the two propagation constants and is usually expressed in terms of the normalized parameter B that is defined by

$$B = \frac{\beta_x - \beta_y}{k} n_{\mathrm{ex}} - n_{\mathrm{ey}} \tag{A1.5.20}$$

where $\beta_x = kn_{\mathrm{ex}}$ and $\beta_y = kn_{\mathrm{ey}}$ in terms of the equivalent effective indices, and k is the free-space wavenumber. If both fundamental mode polarizations are launched simultaneously in the fibre, beating will occur between them because of the difference in propagation constants. The *beat length* or distance over which the superposition of the two modal fields repeats periodically along the length of the fibre is denoted by z_{b},

$$z_{\mathrm{b}} = \frac{2\pi}{\beta_x - \beta_y}; \quad B = \frac{2\pi}{z_{\mathrm{b}} k} = \frac{\lambda}{z_{\mathrm{b}}} \tag{A1.5.21}$$

and is therefore inversely proportional to the birefringence. Hence a measurement of the beat length together with the source wavelength will determine the birefringence.

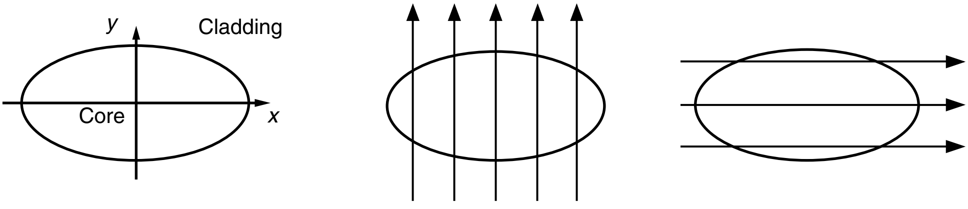


Figure A1.5.10. Cross-section and polarization states of the fundamental mode of an elliptical core fibre.

A1.5.3.12 Attenuation due to absorption and scattering loss

When light propagates in a bound mode along a straight fibre or waveguide, there is no loss of modal power provided that the materials are perfectly loss-less. However, in practical fibres and waveguides, a propagating mode steadily loses power because of two basic physical effects: (i) *bulk absorption* of optical power by the materials constituting the fibre core and cladding; and (ii) *scattering* of light by material and surface inhomogeneities, due principally to the distribution of dopants in the core of fibres and surface roughness between the core and cladding in waveguides.

In long-distance telecommunications fibres, Rayleigh scattering is the major cause of loss, absorption loss having been reduced to almost zero at the operating wavelength. The net effect of these two processes is to reduce the total light in the mode as it propagates along the length of the fibre, i.e. modal power is *attenuated*. The loss of power can be accounted for by adding an imaginary part to the refractive index value, i.e. n becomes the complex index $n^{(\mathrm{r})} + \mathrm{i}n^{(\mathrm{i})}$ where superscripts r and i denote real and imaginary parts, respectively. Although the real part of the index must necessarily vary over the core cross-section to provide guidance, the imaginary part is normally assumed to take a constant value.

Modal attenuation

In the determination of the propagation constant β of each bound mode, a real refractive index distribution $n(x, y)$ leads to a real value of β from the eigenvalue equations as is evident in the modal analyses of the step-profile slab waveguide and fibre in Sections A1.5.4.4 and A1.5.4.5, respectively. If n now becomes complex, it is evident from these eigenvalue equations that the modal parameters U and W and, therefore, the propagation constant must also become complex. Accordingly we set

$$U = U^{(\mathrm{r})} + \mathrm{i}U^{(\mathrm{i})}; \quad W = W^{(\mathrm{r})} + \mathrm{i}W^{(\mathrm{i})}; \quad \beta = \beta^{(\mathrm{r})} + \mathrm{i}\beta^{(\mathrm{i})} \tag{A1.5.22}$$

where $\beta^{(\mathrm{i})} > 0$. If we recall that the electric and magnetic field components all have the common longitudinal dependence

$$\exp(\mathrm{i}\beta z) = \exp(\mathrm{i}\{\beta^{(\mathrm{r})} + \mathrm{i}\beta^{(\mathrm{i})}\}) = \exp(\mathrm{i}\beta^{(\mathrm{r})}z)\exp(-\beta^{(\mathrm{i})}z)$$

it follows that the fields are attenuated as $\exp(-\beta^{(\mathrm{i})}z)$ and the local power density is attenuated as $\exp-(2\beta^{(\mathrm{i})}z)$. Further, since the total power density flow in a mode is the integral of the local power density over the cross-section, it follows that the power in all modes is attenuated according to

$$P(z) = P(0)\exp(-2\beta^{(\mathrm{i})}z) \tag{A1.5.23}$$

where $P(z)$ is the mode power distance z along the fibre or waveguide. The expression $2\beta^{(\mathrm{i})}$ is the *power attenuation coefficient*. The measure of attenuation is commonly expressed in decibels, or dB, where the dB value is calculated as

$$\mathrm{dB} = -10\log_{10}\left(\frac{P(z)}{P(0)}\right) = 20\beta^{(\mathrm{i})}z\log_{10}\mathrm{e} = 8.686\beta^{(\mathrm{i})}z \tag{A1.5.24}$$

The determination of the exact values of the complex propagation constant from the eigenvalue equation for a particular fibre or waveguide with a complex refractive index profile can generally only be undertaken numerically, and formally poses a two-dimensional root-finding problem.

However, for the majority of problems, $\beta^{(\mathrm{i})} \ll \beta^{(\mathrm{r})}$ so that $\beta^{(\mathrm{r})}$ can be determined straightforwardly by assuming that to the lowest order, n is pure real. For example, if the imaginary part of the index $n^{(\mathrm{i})}$ is a constant everywhere throughout the fibre or waveguide, then $\beta^{(\mathrm{i})} \sim kn^{(\mathrm{i})}$ which means that the dB loss is given, approximately, by $55n^{(\mathrm{i})}z/\lambda$ where λ is the source wavelength. This expression is identical to the power attenuation of a plane wave propagating in a uniform medium with imaginary index $n^{(i)}$. For more general situations, a perturbation approach can be adopted to determine an accurate approximation for $\beta^{(i)}$, as quantified by the example in Section A1.5.9.1.1.

For standard silica-based single-mode fibres used for telecommunications, there is a minimum loss of about 0.2 dB km^{-1} occurring at a wavelength of 1550 nm. Using the plane wave expression, this leads to an imaginary index value of approximately 5.6×10^{-12}.

A1.5.3.13 Analytical and numerical solutions

There is only a small number of waveguide geometries and refractive index profiles for which there are exact analytical solutions of Maxwell's equations for the bound mode fields and analytical expressions for the eigenvalue equation that determines the values of the propagation constant. The latter are generally transcendental and can only be solved numerically. Of these solutions, the most practical and simplest example is the step-profile slab waveguide and the step-profile circular fibre. The former has modal fields and eigenvalue equations expressed in terms of trigonometric and exponential functions while the latter has modal fields and eigenvalue equations expressed in terms of Bessel functions and modified Bessel functions (see chapter 12 of Ref. [4]). These solutions are valid for arbitrary relative index difference. In the weak guidance approximation, the corresponding analytical solutions of the scalar wave equation are less complex and are derived in detail in Sections A1.5.4.4 and A1.5.4.5.

Although there are analytical solutions of Maxwell's equations or the scalar wave equation that can be derived for certain other profiles and geometries, they involve special mathematical functions the properties of which are less familiar compared to the situation just 25 years ago. A major factor that has reduced their familiarity is the development and ready availability of reliable commercial and in-house software routines that can solve Maxwell's equations for the fields and propagation constants of bound modes given almost arbitrary refractive index profiles and waveguide geometries.

Planar and rib waveguides

There is a class of waveguides referred to generically as *planar waveguides*. This name actually refers to the planar substrate on which waveguides with a variety of core cross-sections are fabricated. One class relates to *buried channel waveguides* that generally have a nominally square or rectangular core surrounded by an effectively unbounded cladding. Another class includes *rib waveguides* that have a core where the cross-section has the shape of an inverted 'T' and is commonly surrounded by air above and a lower refractive index layer below. Provided the core index is greater than the surrounding cladding index, these waveguides support one or more bound modes, depending on the waveguide parameters and source wavelength.

For both scalar modes and vector modes, the waveguide geometry requires numerical solution for the modal fields and propagation constants, although some analytical approximation methods, such as the *effective index method* can be applied [6].

A1.5.4 Weak-guidance approximation

The solution of Maxwell's equations for the bound modes of a dielectric fibre or waveguide with arbitrary cross-sectional geometry and refractive index profile, whether investigated by analytical or numerical means, generally involves the simultaneous determination of all six scalar components of the electric and magnetic field vectors. Such solutions are therefore necessarily complex. However, there is a set of practical fibre and waveguide problems for which it is possible to make a significant simplification and replace the set of six coupled Maxwell equations with a single scalar equation for just one component of the fields. This is the basis of the *weak guidance approximation* [7, 8].

A1.5.4.1 Scalar electromagnetic fields and power flow

Many waveguides, such as solid silica-based optical fibres and planar waveguides, have a refractive index profile $n(x, y)$ across the core and cladding where the maximum variation in index is relatively small, typically below 1%. Now consider the vector wave equation satisfied by the electric field. This equation is generated by eliminating the magnetic field $\boldsymbol{H}$ between the two Maxwell equations (A1.5.3) and leads to

$$(\nabla_{\mathrm{t}}^2 + k^2 n^2 - \beta^2)E = -\nabla_{\mathrm{t}}(E_{\mathrm{t}} \cdot \nabla_{\mathrm{t}} \ln n^2) \qquad \text{(A1.5.25)}$$

where ∇_{t}^2 is the transverse *vector* Laplace operator, $k = 2\pi/\lambda$ is the free-space wavenumber and β is the propagation constant. If the overall variation in index is small, then the term on the right-hand side can be neglected.

In a general orthogonal coordinate system, the vector operator ∇_{t}^2 couples the scalar components of $\boldsymbol{E}_{\mathrm{t}}$. However, if the components of $\boldsymbol{E}_{\mathrm{t}}$ are chosen to be Cartesian, i.e. $\boldsymbol{E}_{\mathrm{t}} = (e_x, e_y) = \exp(\mathrm{i}\beta z)$ then the component equations decouple. The respective scalar components e_x or e_y for these states independently satisfy the same scalar wave equation

$$(\nabla_{\mathrm{t}}^2 + k^2 n^2 - \beta^2)e_x = 0; \quad (\nabla_{\mathrm{t}}^2 + k^2 n^2 - \beta^2)e_y = 0 \qquad \text{(A1.5.26)}$$

where ∇_{t}^2 is now the two-dimensional *scalar* Laplace operator the analytical form of which depends on the waveguide or fibre geometry.

Once the electric field component of a mode has been determined from the scalar wave equation (A1.5.26), together with the appropriate boundary conditions, the corresponding component of the magnetic field for the respective polarization states is given by a simple algebraic relationship

$$h_y = n_{\mathrm{co}} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} e_x; \quad h_x = -n_{\mathrm{co}} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} e_y \qquad \text{(A1.5.27)}$$

where ε_0 and μ_0 are, respectively, the free-space dielectric constant and permeability, and n_{co} is the maximum core index. Note that in the weak-guidance approximation, all other field components are very small compared to the dominant transverse electric and magnetic field components and can normally be ignored.

It follows from equations (A1.5.9), (A1.5.26) and (A1.5.27) that, in weak guidance, the z-directed power flow density for the x- and y-polarized fields is given, respectively, by

$$S_z = \frac{1}{2} n_{\mathrm{co}} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} e_x^2; \quad S_z = \frac{1}{2} n_{\mathrm{co}} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} e_y^2 \qquad \text{(A1.5.28)}$$

There is a strong analogy between the modes and plane waves of Section A1.5.2 in terms of the non-zero field components and the orthogonal electric field polarization directions. For the two mode polarizations the field components are

$$E = (e_x, 0, 0)\mathrm{e}^{\mathrm{i}\beta z}, H = (0, h_y, 0)\mathrm{e}^{\mathrm{i}\beta z}; \quad E = (0, \mathrm{e}_y, 0)\mathrm{e}^{\mathrm{i}\beta z}, H = (h_x, 0, 0)\mathrm{e}^{\mathrm{i}\beta z} \qquad \text{(A1.5.29)}$$

corresponding *to* x- and y-polarizations, respectively.

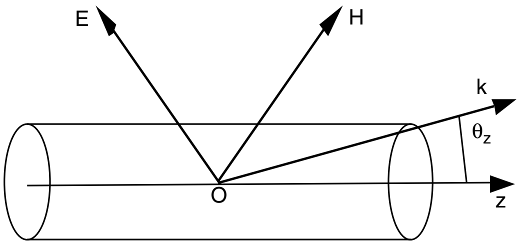


Figure A1.5.11. Triad of electric and magnetic field vectors and the local wave vector.

A1.5.4.2 Transverse nature of the electromagnetic field

The modal electromagnetic field in the core can be considered at each position in the core cross-section as a plane wave. The propagation constant β is the z-component of the local planar wave vector, i.e.

$$\beta = kn_{\mathrm{co}}\cos(\theta_z) \tag{A1.5.30}$$

where θ_z is the direction that the local plane wave vector makes with the z-axis in the core of the waveguide or fibre, and k is the free-space wavenumber. Since the propagation constant β is bounded by

$$kn_{\mathrm{cl}} < \beta < kn_{\mathrm{co}} \tag{A1.5.31}$$

and the core and cladding indices are similar, i.e. $\beta \sim kn_{\mathrm{co}} \sim kn_{\mathrm{cl}}$ leading to $\theta_z \sim 0$.

The wave vector is approximately parallel to the z-direction and hence both the vector electric and magnetic fields are approximately transverse to the z-axis. Both the longitudinal electric and magnetic field components are negligible in magnitude compared to the transverse fields. The Cartesian components of the electric and magnetic fields for either mode polarization, together with the z-axis comprise an orthogonal triad (figure A1.5.11), i.e. the same relationship satisfied by the corresponding components of a plane wave.

A1.5.4.3 Slab and circular geometries

Here the geometry, profiles, field representation, governing equations, coordinates and parameters are delineated for analysing modes on symmetric slab waveguides and circular fibres in the weak guidance approximation.

Symmetric slab waveguide

For the one-dimensional slab waveguide of figure A1.5.12, the core-cladding interfaces are parallel to the y–z plane, so that there is only the x-variation in the transverse direction, together with the usual z-dependent phase term in the z-direction of propagation. Thus for the x- and y-polarizations of the TE and TM modes, respectively, the scalar electric field components can be expressed as

$$E_x(x,z) = e_x(x)\mathrm{e}^{\mathrm{i}\beta z}; \quad E_y(x,z) = e_y(x)\mathrm{e}^{\mathrm{i}\beta z} \tag{A1.5.32}$$

where $e_x(x)$ and $e_y(x)$ are both solutions of the one-dimensional scalar wave equation

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} + k^2n^2(x) - \beta^2\right)e(x) = 0 \tag{A1.5.33}$$

obtained from equation (A1.5.26) where $n(x)$ denotes the profile.

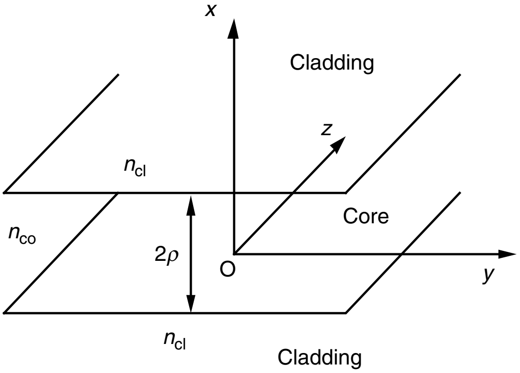


Figure A1.5.12. Geometry and parameters for the symmetric step-profile slab waveguide.

Circular fibre

Note that while either e_x and e_y is the single Cartesian *component* of the vector electric field $\boldsymbol{e}$, their spatial variation does not necessarily have to be described in *Cartesian coordinates*. For the circular fibre in figure A1.5.13 the spatial dependence of the modal fields in the cross-section is best described by polar coordinates (r, ϕ) based on the z-axis of the fibre. Thus for the x- or y-polarized modes we may set

$$e_x = e_x(r, \phi); \quad e_y = e_y(r, \phi) \tag{A1.5.34}$$

Here $e_x(r, \phi)$ and $e_y(r, \phi)$ are both solutions of the two-dimensional scalar wave equation

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2} + k^2 n^2(r) - \beta^2\right) e(r, \phi) = 0 \tag{A1.5.35}$$

where $n(r)$ denotes an axisymmetric profile for circularly symmetric fibres.

Boundary conditions

If ψ denotes either e_x or e_y in the scalar wave equation, then the boundary conditions satisfied by the bound mode solutions of the scalar wave equation in the core and cladding are the weak-guidance limit

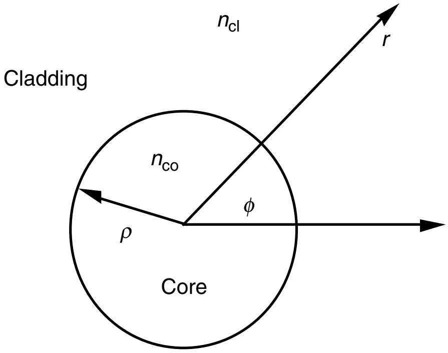


Figure A1.5.13. Geometry and parameters for circular fibres.

of the boundary conditions for Maxwell's equations. In this limit all six components of the electromagnetic field are everywhere continuous. Accordingly:

- $\Psi \rightarrow 0$ exponentially far from the fibre or waveguide core;
- Ψ is continuous across the core–cladding interface; and
- all first derivatives of ψ are continuous across the core–cladding interface.

For the one-dimensional slab waveguide both Ψ and $\mathrm{d}\Psi/\mathrm{d}x$ are continuous across both core–cladding interfaces, while for the two-dimensional fibre Ψ, $\mathrm{d}\Psi/\mathrm{d}r$ and $\mathrm{d}\Psi/\mathrm{d}\phi$ are continuous across the core–cladding interface. The solution of the scalar wave equation together with the boundary conditions constitutes an *eigenvalue problem* for the modal propagation constants, and the *eigenfunctions* determine the spatial dependence of a scalar transverse electric field. The propagation constants are determined by the *eigenvalue equation*.

A1.5.4.4 Step-profile slab waveguide

The spatial dependence of the modal fields in equation (A1.5.32) and the eigenvalue equation for the modal propagation constants are obtained by solving equation (A1.5.33) in the core and the cladding, and then matching the two solutions using the boundary conditions on the core–cladding interfaces. Because the slab waveguide profile is symmetric about the y–z plane, the solutions of the scalar wave equation for $\psi(x)$ will be accordingly either symmetric or anti-symmetric in x.

Even and odd modes

In the core $n(x) = n_{\mathrm{co}}$ and it is convenient to introduce the core modal parameter $U = \rho(k^2 n_{\mathrm{co}}^2 - \beta^2)^{1/2}$ since $\beta < kn_{\mathrm{co}}$ for bound modes, and the normalized coordinate $X = x/\rho$. With these substitutions the scalar wave equation can be written as

$$\left(\frac{\mathrm{d}^2\Psi}{\mathrm{d}X^2} + U^2\right)\Psi = 0 \tag{A1.5.36}$$

This is the harmonic equation and its solutions are proportional to $\cos(UX)$ or $\sin(UX)$. Since the former is symmetric in x, we set for the even modes:

$$\Psi(X) = A\cos(UX) \tag{A1.5.37}$$

for $0 < |X| < 1$, where A is a constant. In the cladding $n(x) = n_{\mathrm{cl}}$ and it is convenient to work with the cladding modal parameter $W = \rho(\beta^2 - k^2 n_{\mathrm{cl}}^2)^{1/2}$ since $\beta > kn_{\mathrm{cl}}$ for bound modes and the normalized $X = x/\rho$. With these substitutions, the scalar wave equation becomes

$$\left(\frac{\mathrm{d}^2\Psi}{\mathrm{d}X^2} + W^2\right)\Psi = 0 \tag{A1.5.38}$$

The solution of this equation is exponential and proportional to either $\exp(WX)$ or $\exp(-WX)$. Since we require a symmetric solution that decreases to zero as $X \rightarrow +\infty$ and $X \rightarrow -\infty$, we set

$$\Psi(X) = B\,\mathrm{e}^{-W|X|} \tag{A1.5.39}$$

for $|X| > 1$, where B is a constant.

Eigenvalue equations

The boundary conditions on ($X = 1$) ($x = \rho$) and ($X = -1$) ($x = -\rho$), which link the core and cladding solutions, are the continuity of Ψ and $\mathrm{d}\Psi/\mathrm{d}X$, which lead, respectively, to

$$A\cos U = B\,\mathrm{e}^{-W}; \quad -AU\sin U = -BW\,\mathrm{e}^{-W} \tag{A1.5.40}$$

so that on dividing the two equations, the constants A and B are eliminated and we obtain one equation linking the core and cladding modal parameters U and W. A second equation follows from the definitions of U, V and W. Hence the eigenvalue equations for the even modes are

$$W = U\tan U; \quad W^2 + U^2 = V^2 \tag{A1.5.41}$$

The derivation of the corresponding equations for the odd modes is identical to that for the even modes provided the core modal field has the anti-symmetric form

$$\Psi(X) = A\sin(UX) \tag{A1.5.42}$$

This leads to the eigenvalue equations for the odd modes

$$W = -U\cot U; \quad W^2 + U^2 = V^2 \tag{A1.5.43}$$

Both pairs of eigenvalue equations can be further simplified by eliminating U between each pair and for the even and odd modes this leads, respectively, to

$$V = \pm\frac{U}{\cos U}; \quad V = \pm\frac{U}{\sin U} \tag{A1.5.44}$$

but care needs to be exercised with the signs to identify correct solutions.

Solution of the eigenvalue equations

The eigenvalue equations for both even and odd modes are transcendental and do not possess closed-form analytical solutions for U (or W) in terms of a given value of V. Accordingly, solutions must be determined numerically and each value of U corresponds to one bound mode. This can be undertaken using a hand calculator incorporating trigonometric functions by using trial and error, i.e. given V guess U, or by using a more sophisticated computer program.

However, it may be simpler to calculate the value of V, given the value of U in equations (A1.5.44) since these are single-valued and explicit. This approach generates the V–U curve for each and every mode of the waveguide as shown in figure A1.5.14. In these plots, an even or odd subscript corresponds to a mode with an even or odd field, respectively, and also indicates mode order.

Single-mode waveguide

In figure A1.5.14 every mode, except the fundamental mode, has a finite cut-off value of U when $U = V$ below which it cannot propagate as a bound mode and becomes a leaky mode. For the mode with subscript ‘m’, the cut-off value is $V = V_{\mathrm{co}} = m\pi/2$.

The waveguide is *single mode* if $V < \pi/2$, and only the fundamental mode can propagate in either of its two polarization states (TE_0 or TM_0) with the same propagation constant β. The fundamental mode has the largest value of propagation constant or equivalently the smallest value of U for a given value of V. The fundamental mode corresponds to the TE_0 or TM_0 mode.

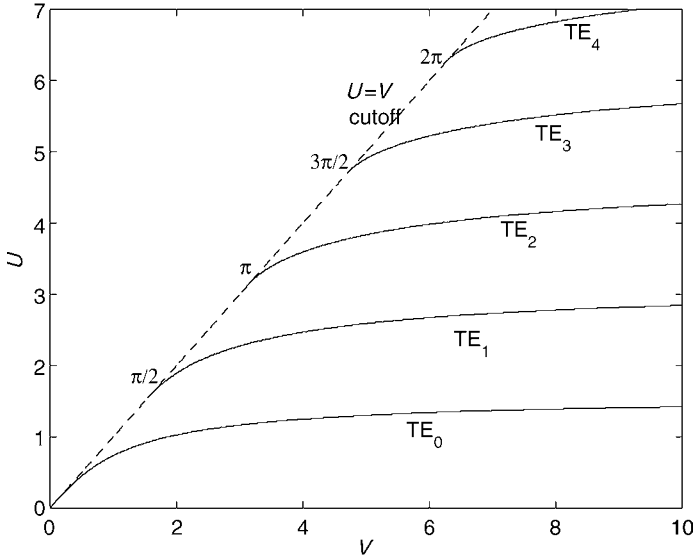


Figure A1.5.14. Plots of U against V for the modes of the symmetric step-profile slab waveguide.

Modal fields

For the TE (transverse electric) modes the electric field direction is parallel to the y-axis and transverse to the to the z-axis while for the TM (transverse magnetic) modes, the magnetic field is parallel to the y-axis and transverse to the z-axis. In the weak-guidance approximation, the pair of TE$_j$ and TM$_j$ modes has an identical propagation constant while their transverse electric and magnetic field components have identical spatial distribution, and are y- and x-polarized, respectively.

The spatial distribution of the transverse electric or magnetic fields for the first four TE or TM modes are plotted in figure A1.5.15. Note that the TE$_m$ and TM$_m$ mode fields have $m+1$ extrema, and the even mode has a maximum at the centre of the waveguide where the odd mode fields vanish. The total number of modes that can propagate for a given value of V is given by the nearest integer value below $4V/\pi$ allowing for the two possible polarization states.

A1.5.4.5 Step-profile fibre

The spatial dependence of the modal fields in equation (A1.5.34) and the eigenvalue equation for the modal propagation constants are obtained by solving equation (A1.5.35) in the core and the cladding, and then matching the two solutions using the boundary conditions on the core–cladding interface. If $\Psi(r, \phi)$ denotes either the x-polarized electric field $e_x(r, \phi)$ or the y-polarized electric field $e_y(r, \phi)$ in cylindrical polar coordinates, then, within weak guidance, $\Psi(r, \phi)$ satisfies the scalar wave equation (A1.5.34), which is more conveniently written in the normalized form

$$\left\{\frac{\partial^2}{\partial R^2}+\frac{1}{R}\frac{\partial}{\partial R}+\frac{1}{R^2}\frac{\partial^2}{\partial \phi^2}+\rho^2(k^2n(R)^2-\beta^2)\right\}\Psi=0 \qquad \text{(A1.5.45)}$$

where $R=r/\rho$ is the normalized radial coordinate, where $n(R)$ is the refractive index profile and ρ is the core radius. This equation is a second-order partial differential equation and as $n(R)$ is independent of ϕ it can be solved using *separation of variables*.

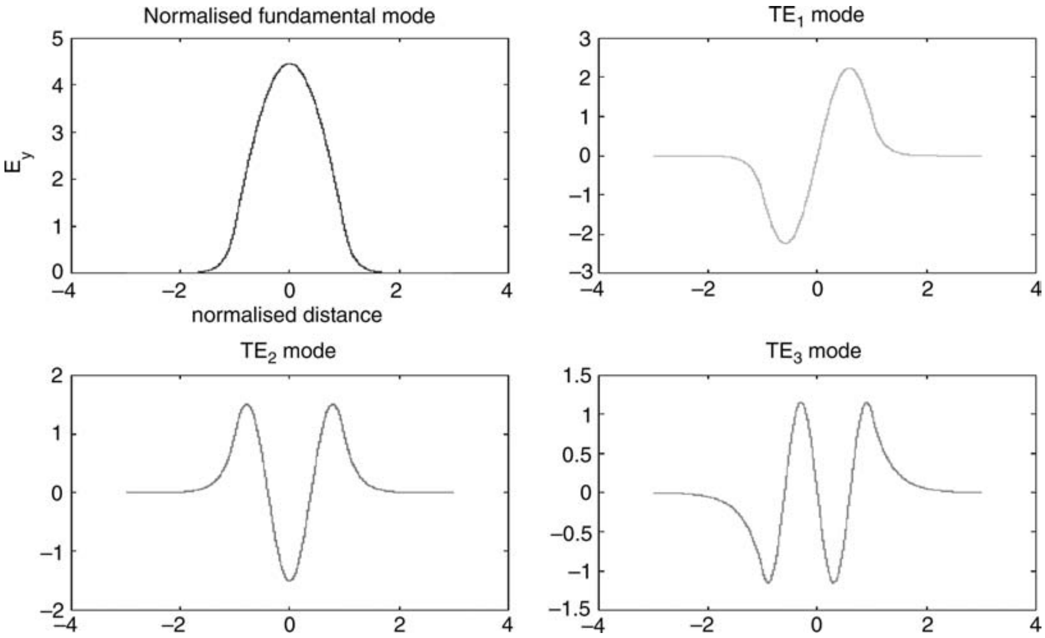


Figure A1.5.15. Plots of the transverse electric field for the first four modes of the symmetric step-profile slab waveguide.

Separation of variables

Using this technique, the two-dimensional spatial dependence of ψ is expressed as the product of the two single-variable functions

$$\Psi(R, \phi) = F(R)G(\phi) \tag{A1.5.46}$$

where $F(R)$ is a function of R only, and $G(\phi)$ is a function of ϕ only. On substituting into the scalar wave equation (A1.5.38), dividing by Ψ, and multiplying by R^2 it follows

$$\frac{R^2}{F}\left\{\frac{\mathrm{d}^2F}{\mathrm{d}R^2} + \frac{1}{R} - \frac{\mathrm{d}F}{\mathrm{d}R} + \rho^2\left[k^2n^2(R) - \beta^2\right]F\right\} + \frac{1}{G}\frac{\mathrm{d}^2G}{\mathrm{d}\phi^2} = 0 \tag{A1.5.47}$$

The right-hand expression is a function of ϕ only and can be held fixed while R varies arbitrarily. Because of the periodicity of the solution required in the azimuthal direction because of the circular geometry we set

$$\frac{\mathrm{d}^2G}{\mathrm{d}\phi^2} = -\nu^2 G; \quad G = \begin{cases} \sin(\nu\phi) \\ \cos(\nu\phi) \end{cases} \tag{A1.5.48}$$

where ν is a constant. This is the harmonic equation with sinusoidal solutions which are single-valued, provided $\nu = 0, 1, 2, 3, \ldots$. Substituting back into the scalar wave equation (A1.5.38) leads to the second-order ordinary differential equation

$$\left\{\frac{\mathrm{d}^2}{\mathrm{d}R^2}+\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}R}-\frac{\nu^2}{R^2}+\rho^2[k^2n^2(R)-\beta^2]\right\}F=0 \qquad \text{(A1.5.49)}$$

that is solved for the core and cladding regions separately.

Core solution

In the core $n = n_{\mathrm{co}}$, so that equation (A1.5.49) for F can be written as

$$\left\{\frac{\mathrm{d}^2}{\mathrm{d}R^2}+\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}R}-\frac{\nu^2}{R^2}+U^2\right\}F=0 \qquad \text{(A1.5.50)}$$

where the core modal parameter is defined by $U=\rho(k^2n_{\mathrm{co}}^2-\beta^2)^{1/2}$. Multiply this equation by R^2 and set $s = UR$ to obtain

$$\left\{s^2\frac{\mathrm{d}^2}{\mathrm{d}s^2}+s\frac{\mathrm{d}}{\mathrm{d}s}+s^2-\nu^2\right\}F=0 \qquad \text{(A1.5.51)}$$

This is *Bessel's equation* with the general solution

$$F(R)=AJ_\nu(s)+BY_\nu(s) \qquad \text{(A1.5.52)}$$

where J_ν and Y_ν are *Bessel functions* of the first and second kinds, respectively, and A and B are constants. The solution is bounded throughout the core and as Y_ν is singular on the fibre axis $R = 0$, then $B = 0$. Hence $F(R) = AJ_\nu(UR)$ and together with equation (A1.5.48) the complete core solution dependence $F(R)G(\phi)$ is given by

$$\Psi=AJ_\nu(UR)\begin{cases}\sin(\nu\phi)\\ \cos(\nu\phi)\end{cases} \qquad \text{(A1.5.53)}$$

where $\nu = 0, 1, 2, \ldots.$

Cladding solution

In the cladding $n = n_{\mathrm{cl}}$, so that the corresponding equation for F becomes

$$\left\{s^2\frac{\mathrm{d}^2}{\mathrm{d}s^2}+s\frac{\mathrm{d}}{\mathrm{d}s}-(\nu^2+s^2)\right\}F=0 \qquad \text{(A1.5.54)}$$

where $s = WR$ and the cladding modal parameter is defined by $W=\rho(\beta^2-k^2n_{\mathrm{cl}}^2)^{1/2}$. This is the *modified Bessel equation* with the general solution

$$F(R)=CI_\nu(s)+DK_\nu(s) \qquad \text{(A1.5.55)}$$

where I_ν and K_ν are *modified Bessel functions* of the first and second kinds, respectively, and C and D are constants. This solution is bounded throughout the cladding and since I_ν is singular as $R \to \infty$, then $C = 0$. Hence $F(R) = DK_\nu(WR)$ and together with equation (A1.5.48) the complete cladding solution dependence, $F(R)G(\phi)$ is

$$\Psi=DK_\nu(WR)\begin{cases}\sin(\nu\phi)\\ \cos(\nu\phi)\end{cases} \qquad \text{(A1.5.56)}$$

where $\nu = 0, 1, 2, \ldots.$

Boundary conditions

Boundary conditions at the core–cladding interface require continuity of the solution of the scalar wave equation and all first derivatives, equivalent to continuity of $\Psi(R,\phi)$ and $\mathrm{d}\Psi(R,\phi)/\mathrm{d}R$ on $R = 1$. The latter, when applied to equations (A1.5.53) and (A1.5.56), gives

$$AJ_\nu(U) = DK_\nu(W); \quad AUJ'_\nu(U) = DWK'_\nu(W) \tag{A1.5.57}$$

where the prime denotes differentiation with respect to the argument. On dividing these two equations, we obtain one eigenvalue equation for U and W

$$\frac{UJ'_\nu(U)}{J_\nu(U)} = W\frac{K'_\nu(W)}{K_\nu(W)} \tag{A1.5.58}$$

It is more convenient for computational purposes to recast this equation into an alternative form avoiding the use of derivatives by using the following recurrence relations satisfied by Bessel functions of different orders

$$J'_\nu = \frac{\nu}{U}J_\nu - J_{\nu+1}; \quad K'_\nu = \frac{\nu}{W}K_\nu - K_{\nu+1} \tag{A1.5.59}$$

where the prime denotes differentiation with respect to the argument. Thus, the final form of the eigenvalue equations for U and W are

$$U\frac{J_{\nu+1}(U)}{J_\nu(U)} = W\frac{K_{\nu+1}(W)}{K_\nu(W)}; \quad V^2 = U^2 + W^2 \tag{A1.5.60}$$

where $\nu = 0, 1, 2, \ldots.$

Solutions of the eigenvalue equations

The left-hand term in equation (A1.5.60) is transcendental and can be solved only numerically. There are standard library routines for evaluating the two types of Bessel functions required. Given the value of V, the eigenvalue equations can be solved by expressing W in terms of U and V using the second term in equation (A1.5.60) and then solving the resulting equation numerically for the discrete values of U. Each value of U corresponds to a particular bound mode. For each value of ν, there are a finite number of bound solutions of the eigenvalue equation, the number increasing with the value of V according to the closest integer below $V^2/2$ for the step profile.

Figure A1.5.16 plots the values of U as a function of V for each bound mode solution. Given the value of U, the value of the propagation constant can be determined from the definition of U given above. Every mode, except the fundamental mode, has a finite cut-off value of U when $U = V$. Below this value the mode cannot propagate as a bound mode and becomes a leaky mode. The fibre is multimode when $V \gg 1$, and many modes can propagate. For a given, large value of V, the number of bound modes which can propagate is approximately $V^2/2$.

Mode nomenclature

Because of the cylindrical geometry of the fibre, there are four possible types of bound mode solution. The TE_{01} and TM_{01} solutions, like the corresponding solutions for the slab waveguide in Section A1.5.4.4, indicate modal electric or magnetic fields that have only a radial electric or magnetic field component, respectively, and are axisymmetric. The EH_{pq} and HE_{pq} modes can be regarded as hybrid modes, their fields containing both electric and magnetic radial components. The first and second

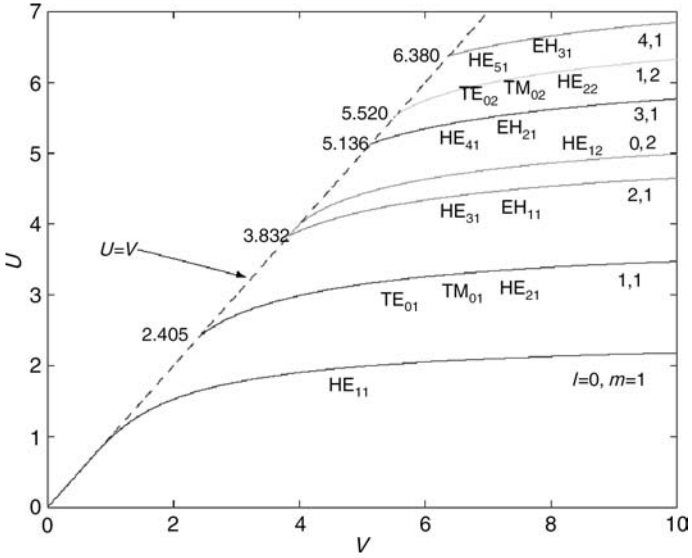


Figure A1.5.16. Plots of U against V for modes of the step-profile circular fibre.

subscripts p and q on each mode designator indicate, respectively, the azimuthal symmetry and radial order. The former has the dependence $\cos([p-1]\phi)$ or $\sin([p-1]\phi)$ for $p = 1, 2, 3 \ldots$, while the latter indicates the number of extrema in the field profile.

There is a second mode nomenclature in use that was devised specifically for weakly guiding fibres and is based on the linear polarization (LP) property of these modes. In this system the fundamental HE_{11} mode is known as the LP_{01} mode with axisymmetric field components, and the TE_{01}, TM_{01} and HE_{21} set of second modes with anti-symmetric $\sin\phi$ or $\cos\phi$ azimuthal dependence as LP_{11} modes.

Single-mode fibre

The fibre is single-mode for $V < 2.405$ when only one mode, the fundamental mode, can propagate in either of its two orthogonal but otherwise arbitrary polarization states with the same propagation constant β. The fundamental mode has the largest value of propagation constant or equivalently the smallest value of U for a given value of V. The fundamental mode corresponds to the HE_{11} or equivalent LP_{01} mode.

Fractional of modal power in the core

For a particular fibre and source wavelength, the power of each propagating mode is distributed between the core and the cladding in a ratio that depends on the mode order. It is useful to define the parameter η as the fraction of total mode power that propagates in the core, so that $0 < \eta < 1$. Figure A1.5.17 plots this fraction as a function of fibre parameter V for the first six low-order modes of the weakly guiding step profile. As V increases the value of η for each mode approaches 1 asymptotically, consistent with the zero-wavelength geometric optics limit. Note that as V approaches the mode cut-off, the fraction of core power does not always approach zero.

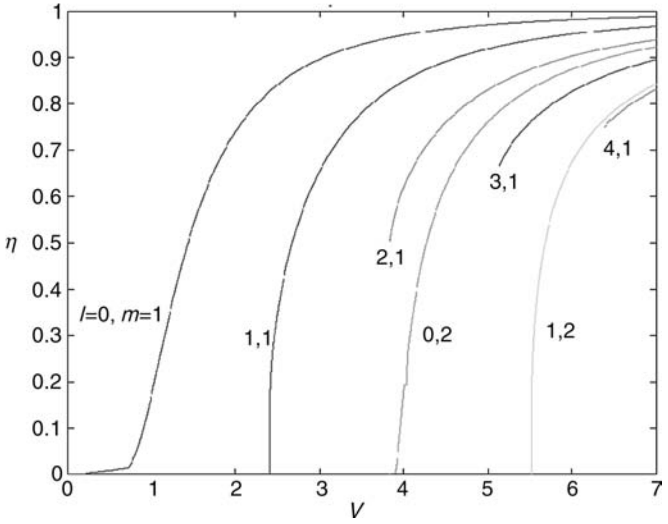


Figure A1.5.17. Fraction of modal power propagating in the fibre core.

A1.5.4.6 Fundamental mode

The mode of most interest for single-mode fibres is the fundamental HE_{11} or LP_{01} mode, i.e. the only mode that can propagate on a fibre for V values below the cut off of the second mode, namely for $V < 2.405$ in the case of the step-profile fibre. If λ_{co} denotes the cut-off wavelength of the second mode, the fundamental mode alone can propagate for wavelengths such that $\lambda > \lambda_{co}$.

Intensity distribution

The normalized fundamental mode electric field intensity $|\Psi|^2$ for typical graded- or step-index profiles has a characteristically axially symmetric bell-shaped distribution about the fibre axis. This characteristic shape is exemplified by the plots of intensity against normalized radius, $R = r/\rho$, shown in figure A1.5.18 calculated from equations (A1.5.28), (A1.5.53) and (A1.5.56) for the step-profile fibre for various values of V. In the geometric optics limit when V becomes unbounded, i.e. $\lambda \rightarrow 0$, all the modal power is concentrated in the core. The normalization is such that the geometric optics limit corresponds to unit intensity on the fibre axis.

Fibre cladding thickness

As V decreases, an increasing fraction of fundamental mode power spreads into the cladding, as is clear from figures A1.5.17 and A1.5.18. For a practical single-mode step-profile fibre operating around $V = 2.3$, the fraction of modal power in the cladding is about 25% and the cladding field, which decreases approximately exponentially with increasing radius, needs to be sufficiently small at the outer edge of the cladding so that the absorption due to the fibre coating is negligibly small. It is for this reason that a standard 125 μm cladding diameter has been adopted resulting in a ratio of cladding-to-core diameters in the range of 10–15, depending on core size.

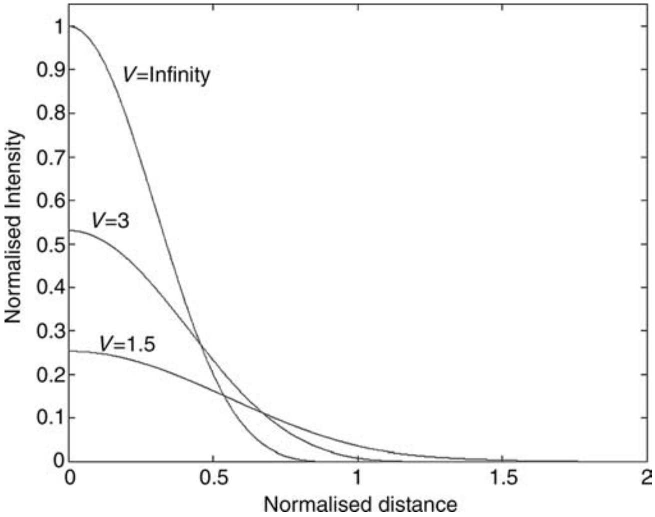


Figure A1.5.18. Plots of the fundamental mode intensity distribution on a step-profile fibre for various values of *V*.

A1.5.4.7 Spot size and Gaussian approximation

Single-mode fibres and waveguides are often characterized by the *spot size* of the fundamental mode intensity distribution, which is an experimental measurement of the diameter of the brightest part of this distribution as it appears in either the near field or far field at the end of a long length of single-mode fibre. The plots of the local power density of the fundamental mode transverse intensity distribution, as shown in figure A1.5.18, have the common characteristic of a peak on the vertical axis and a monotonic decrease with increasing radius. Accordingly, for a given wavelength, the *spot size* can be defined as the half-width of this distribution at an appropriate position on the vertical axis while the *mode field diameter* is the full width and therefore equal to twice the spot size.

As there is no immediately obvious position to choose, a number of different definitions of spot size have been proposed over the years, each being tailored to provide, indirectly, an accurate value of a particular fundamental mode property from the spot size measurement. Here we introduce a definition known as the *Gaussian approximation* based solely on the shape of the plots in figure A1.5.18 (see chapters 15 and 16 of Ref. [4]).

Gaussian approximation

Generally there is no analytical solution of the scalar wave equation for the fundamental mode field of a practical profile, but this equation can of course always be solved numerically. However, a simple but approximate analytical representation for this field can be derived by examining the intensity distribution of the fundamental mode in figure A1.5.18. At a sufficiently large normalized radial distance R from the fibre axis, the asymptotic form of the square of the modified Bessel function $K_0(WR)$ intensity decreases approximately exponentially as $\exp(-2WR)$, where W is the cladding mode parameter. Furthermore, the intensity also exhibits an approximately parabolic variation with R around its maximum on the fibre axis. This is evident for a general index profile by examining a power series solution of equation A1.5.51, and in the case of the step-profile fibre the analytical dependence of the square of the core field on $J_0(UR)$ in equation A1.5.53 exhibits this property for small values of the argument of the Bessel function.

A standard mathematical function that reflects these attributes and is algebraically tractable for integration in particular is the *Gaussian function*. It is used as an approximate representation of the fundamental mode field according to

$$\Psi(R) = A\exp\left(\frac{1}{2}\frac{r^2}{s^2}\right) = A\exp\left(\frac{1}{2}\frac{R^2}{S^2}\right) \tag{A1.5.61}$$

where A is an arbitrary constant, s is the spot size and $S = s/\rho$ is the normalized spot size. The factor of 1/2 is purely for convenience in subsequent algebraic manipulations. This function has an approximately parabolic variation in the neighbourhood of the fibre axis. In the far field at $R \gg 1$, the function has an exponential decrease that varies as R^2 rather than linearly but since the fraction of the modal power in this region is small, the additional error will be small.

The normalized spot size S is determined from the fibre profile, using standard variational methods to derive the following implicit equation (see section 15.1 of Ref. [4])

$$\frac{1}{V^2} = \int_0^\infty R^2\frac{\mathrm{d}f(R)}{\mathrm{d}R}\exp\left(-\frac{R^2}{S^2}\right)\mathrm{d}R; \quad n^2(R) = n_{\mathrm{co}}^2\{1 - 2\Delta f(r)\} \tag{A1.5.62}$$

where $f(R)$ describes the profile variation, V is the fibre parameter and Δ the relative index difference. For the step profile and the Gaussian profile with $f(R) = 1 - \exp(-R^2)$ over $0 < R < \infty$, this leads respectively to

$$S = \frac{1}{2\ln V^{1/2}}; \quad S = \frac{1}{(V-1)^{1/2}} \tag{A1.5.63}$$

with the limitation that $V > 0$ for the former and $V > 1$ for the latter. These expressions can then be used to generate analytical expressions for other modal properties of interest.

A1.5.4.8 Depressed-cladding and W-fibres, fundamental mode cut-off

The fibres and waveguides studied so far have a uniform cladding index and a core refractive index profile that is either uniform or graded but has a minimum index equal to or greater than the cladding. This type of profile is known as a *matched-cladding* profile.

Depressed-cladding fibres

By contrast, a *depressed-cladding* fibre has a depressed ring-shaped region between the core and cladding that is relatively wide compared to the core radius, as shown by the left-hand profile in figure A1.5.19. The term arises from the index miss-match that occurs between the deposited silica and the silica of the

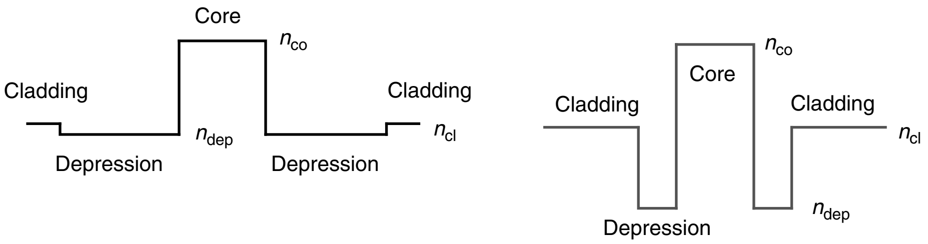


Figure A1.5.19. Schematics for depressed-cladding and W profiles.

starting preform tube used in the fibre fabrication process. The depressed ring has a slightly lower index n_{dep} compared with that of the cladding.

W-fibres

A second type of fibre with a depressed region is known as a *W-fibre*. It is characterized by a depression that is relatively narrow compared to the core radius, but has a much lower index n_{dep} compared with that of the depressed-cladding fibre, as indicated by the right-hand profile in figure A1.5.19. W-fibres were originally developed as they offer better dispersion characteristics than a matched-cladding fibre, in keeping with the discussion of pulse dispersion and multi-layered fibres for dispersion flattening in Section A1.5.6.4.

Fundamental mode cut-off

W-fibres and depressed-cladding fibres are normally designed to be single mode relative to the core–cladding index difference for a particular wavelength. Thus, they can support only the fundamental bound mode with an effective index value n_{eff} such that $n_{\mathrm{cl}} < n_{\mathrm{eff}} < n_{\mathrm{co}}$. On matched-cladding fibres the fundamental mode does not have a finite cut-off wavelength below which it cannot propagate as a bound mode, whereas on depressed-cladding and W-fibres the fundamental mode may have a finite cut-off wavelength depending on the width and depth of the depressed region.

There is a simple algebraic condition that determines whether or not the fundamental mode has a finite cut-off wavelength on a depressed-cladding or W-fibre. This condition may be expressed in terms of the following integrals for circular fibres and slab waveguides, respectively

$$I = \int_0^{\infty} \left\{ n^2(r) - n_{\mathrm{cl}}^2 \right\} r \, \mathrm{d}r; \quad I = \int_{-\infty}^{\infty} \left\{ n^2(x) - n_{\mathrm{cl}}^2 \right\} \mathrm{d}x \tag{A1.5.64}$$

where n is the complete refractive index profile and n_{cl} is the uniform cladding index. If $I > 0$, the fundamental mode does not have a finite cut-off wavelength, but if $I < 0$, there is a finite cut-off wavelength. Note that this condition does not determine the value of the cut-off wavelength; this needs to be evaluated from the appropriate eigenvalue equation by setting $\beta = kn_{\mathrm{cl}}$ and determining the corresponding wavelength. This result has obvious applications in the design of fibres.

A1.5.5 Ray tracing

The propagation of guided light along any fibre or waveguide is governed by Maxwell's equations and can be formally analysed exactly in terms of a superposition of bound modes as described in Sections A1.5.3 and A1.5.4. However, when there are a large number of modes present, as occurs in multimode waveguides and fibres, there is an alternative and accurate approach to the analysis of propagation using *ray tracing*.

Ray tracing is based on the zero-wavelength limit of Maxwell's equations and assumes: that (a) propagation is governed only by the local value of the refractive index and its variation within the core and (b) it provides a continuum solution as opposed to a summed discretized modal solution.

For real sources, the wavelengths used in optoelectronics are relatively small but non-zero, and where there are many modes present, the modal superposition and ray solutions are almost equivalent, the relative error decreasing with increasing values of V. For these situations, rays can be regarded as local plane waves. One advantage of ray tracing is that in some situations it provides a more physical description and interpretation of propagation compared to a summed modal analysis.

A1.5.5.1 Snell's laws

For refractive index distributions that are piece-wise continuous, propagation in the uniform media away from interfaces is described by straight-line ray trajectories equivalent to plane-wave propagation. At interfaces between regions of uniform but different indices, the transmission, reflection and refraction of incident rays is described by *Snell's laws*.

Consider the planar interface shown in figure A1.5.20 between uniform media of refractive indices n_{co} and n_{cl} where $n_{\mathrm{co}} > n_{\mathrm{cl}}$. A ray is incident in the medium of index n_{co} at angle θ_z relative to the interface. If $\theta_z > \theta_{\mathrm{c}}$ where θ_{c} is the *complementary critical angle* relative to the interface (rather than the normal) and is defined by

$$\sin(\theta_{\mathrm{c}}) = \left(1 - \frac{n_{\mathrm{cl}}^2}{n_{\mathrm{co}}^2}\right)^{1/2}; \quad \cos(\theta_{\mathrm{c}}) = \frac{n_{\mathrm{cl}}}{n_{\mathrm{co}}} \tag{A1.5.65}$$

then the incident ray is known as a *refracting ray* and is both reflected and refracted as shown in figure A1.5.20(*a*). The angle of reflection is equal to the angle of incidence and the angle of transmission θ_{t} satisfies $n_{\mathrm{co}}\cos\theta_z = n_{\mathrm{cl}}\cos\theta_{\mathrm{t}}$. Since $n_{\mathrm{co}} > n_{\mathrm{cl}}$, the angle of transmission is smaller than the angle of incidence. An important consequence of refraction is that only a fraction of the incident ray power is reflected and the balance is transmitted.

If $\theta_z < \theta_{\mathrm{c}}$ then the incident ray is a *reflecting ray* and is only reflected as shown in figure A1.5.20(*b*), where the angle of reflection is equal to the angle of incidence. In this situation, all the power in the incident ray enters the reflected ray.

In the weak guidance approximation, n_{co} is only slightly larger than n_{cl}, so that the complementary critical angle is relatively small and only rays incident at small angles relative to the interface are totally reflected.

A1.5.5.2 Eikonal equation

In graded media where the refractive index varies continuously with position, the ray paths are no longer straight lines and a generalized form of Snell's laws is required to determine the curved ray trajectories. Inasmuch that Snell's laws can be derived using plane-wave incidence, reflection and transmission across the interface in figure A1.5.20, the governing *eikonal equation* for curved ray paths can be derived from Maxwell's equations for a graded index distribution in the limit of zero wavelength [9]. In equation (A1.5.66) the first expression gives this equation in general three-dimensional form where $\boldsymbol{r}(x, y, z)$ is the position vector, $n(\boldsymbol{r})$ is the refractive index distribution and s is the distance along the curved path of the ray. The second and third expressions apply to a slab waveguide with a transverse variation of index $n(x)$.

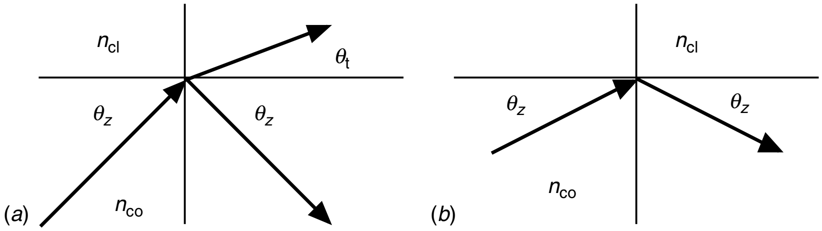


Figure A1.5.20. Snell's laws at the core–cladding interface.

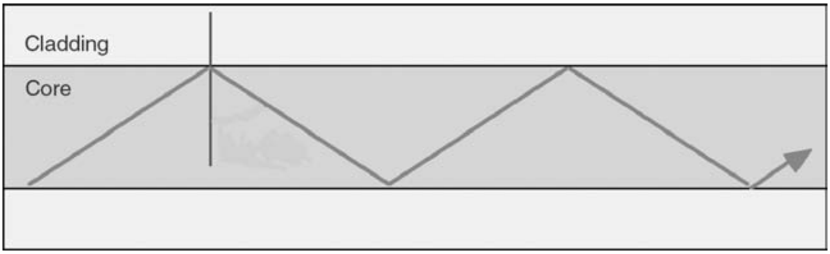


Figure A1.5.21. Zig-zag bound ray paths on a step-profile slab waveguide.

$$\frac{\mathrm{d}}{\mathrm{d}s}\{n(r)\}\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s} = \nabla n(r); \quad \frac{\mathrm{d}}{\mathrm{d}s}\{n(x)\}\frac{\mathrm{d}x}{\mathrm{d}s} = \frac{\mathrm{d}n(x)}{\mathrm{d}s}; \quad \frac{\mathrm{d}}{\mathrm{d}s}\{n(x)\}\frac{\mathrm{d}z}{\mathrm{d}s} = 0 \qquad \text{(A1.5.66)}$$

Similar forms apply to circular fibres.

A1.5.5.3 Step-profile multimode slab waveguide

Assuming a symmetric, step-profile slab waveguide with a core of uniform refractive index n_{co} surrounded on either side by a cladding of uniform but lower index n_{cl}, there are two types of ray paths that can propagate along the waveguide.

Bound ray paths

These are zig-zag, straight-line paths that Snell's laws predict are totally reflected from alternate core–cladding interfaces (figure A1.5.21) at a constant angle θ_z relative to the z-direction provided that this angle is smaller than the complementary critical angle, i.e. $\theta_z < \theta_{\mathrm{c}}$. If the waveguide materials are lossless, bound rays can propagate indefinitely along the straight waveguide and hence the total power carried by all bound rays must remain constant everywhere along the length of the waveguide.

Refracting ray paths

These are zig-zag, straight-line paths that are both reflected and refracted at successive core–cladding interfaces (figure A1.5.22) and make a constant angle θ_z relative to the z-direction, such that $\theta_z > \theta_{\mathrm{c}}$. However, the effect of refraction is to remove a constant fraction of power from the ray propagating within the waveguide at each reflection, which is equivalent to an approximately exponential decrease in power with distance z along the waveguide. Accordingly, refracted rays can be referred to as *leaky rays* as their power decreases along the length of the waveguide, i.e. they are attenuated.

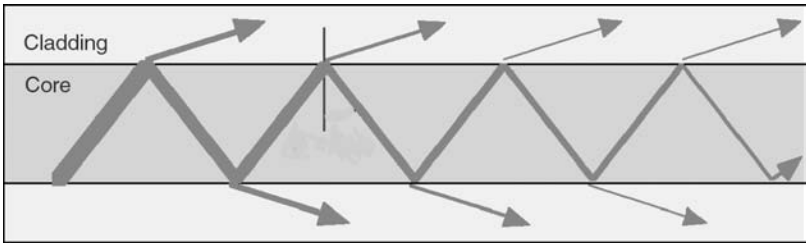


Figure A1.5.22. Zig-zag refracting ray paths on a step-profile slab waveguide.

Refracting rays with a propagation angle θ_z that is not too close to the complementary critical angle generally attenuate very quickly. This means that after a short distance along the waveguide, virtually all their power is lost to the cladding. Rays with θ_z close to the complementary critical angle attenuate more slowly and persist farther along the waveguide. After a sufficient distance essentially only the power in bound rays remains.

A1.5.5.4 Graded-profile multimode slab waveguide

Given a particular graded-index profile, the ray paths along a slab waveguide are determined from the solution of the two-dimensional form of the eikonal equation (A1.5.66). In regions where the index is decreasing away from the index, ray paths propagating away from the waveguide axis curve back towards the waveguide axis and conversely ray paths propagating towards the axis curve away from the axis. Put together, this means that bound rays follow a quasi-sinusoidal path that is periodic along the waveguide.

The distance over which the path repeats itself is known as the *ray period*, and the extremity of each path touches a line on either side of the axis, known as the *ray caustic*. This caustic and its counterpart on the opposite side of the waveguide z-axis constitute the bounding domain for all ray paths with the same angle relative to the z-axis regardless of the longitudinal starting position. In general, the ray period will vary with the angle that the ray path makes with the waveguide axis. However, there is a profile for which all ray directions have exactly the same ray period.

Hyperbolic secant profile

The hyperbolic secant profile is symmetric about the waveguide axis and is defined by

$$n^2(x) = n_{\mathrm{co}}^2 \sec^2(\kappa x) \tag{A1.5.67}$$

where x is the transverse coordinate, n_{co} is the maximum index on the waveguide axis and κ is a scaling constant. The two-dimensional eikonal equations (A1.5.66) can be solved analytically for this profile and confirm that every ray path has the *same ray period* z_{p} regardless of the angle it makes with the z-axis where $z_{\mathrm{p}} = \pi/\kappa$ (figure A1.5.23). Such a waveguide is sometimes referred to as a GRIN (for gradient-index) waveguide.

These waveguides also have the property that the ray half-period $z_{\mathrm{p}} = \pi/\kappa$ is independent of wavelength. Furthermore, the hyperbolic secant profile produces exactly zero inter-modal dispersion in

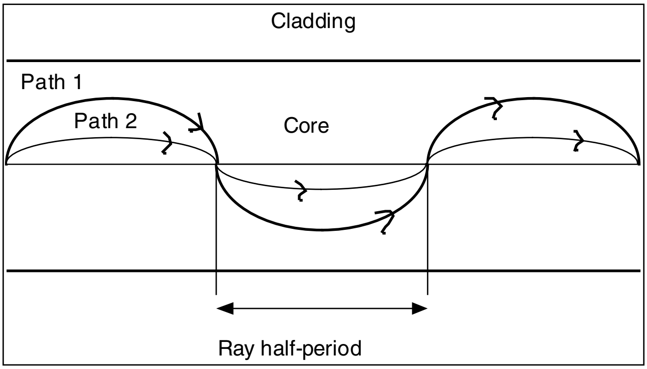


Figure A1.5.23. Bound ray paths with a common ray period on the hyperbolic-profile slab waveguide.

the slab waveguide as discussed in Section A1.5.6.5, so that waveguide and material dispersion may need to be taken into account.

A1.5.5.5 Multimode fibres

Ray tracing in step- and graded-profile multimode fibres comprises a straightforward generalization of the ray path tracing techniques for slab waveguides from two to three dimensions, but the reflection of straight-line core rays from the curved core–cladding interface of a step-profile fibre depends on wavelength. In the limit of zero wavelength, rays are either reflected or reflected/refracted depending on the angle the ray path makes with the tangential plane on the interface, independent of the curvature of the interface. However, if the wavelength is non-zero, a third class of ray paths, known as *tunnelling rays*, appears.

Tunnelling rays

If a ray, or local plane wave with a small but finite wavelength, is incident on the curved interface between the core and cladding of a step-profile fibre, it will be refracted if the angle of incidence α_i it makes relative to the normal at the reflection point is less than the classical critical angle, i.e. $\alpha_\mathrm{i} < \sin^{-1}(n_\mathrm{cl}/n_\mathrm{co})$. For $\alpha_\mathrm{i} < \sin^{-1}(n_\mathrm{cl}/n_\mathrm{co})$ there are two possible classifications of rays, depending on the angle the ray direction makes with the z-axis of the fibre, i.e. θ_z. If $\theta_z < \theta_\mathrm{c}$ where θ_c is the complementary critical angle, then the ray is totally internally reflected, but if $\theta_z < \theta_\mathrm{c}$ such that $\alpha_\mathrm{i} < \sin^{-1}(n_\mathrm{cl}/n_\mathrm{co})$ then the ray is partially reflected and transmitted and is known as a *tunnelling ray*. Accordingly, refracting and tunnelling rays are commonly referred to as *leaky rays*, since they lose power at each reflection when propagating along the core of the fibre.

The delineation between the three classes of ray is depicted in figure A1.5.24, where the dark half-cone denotes refracting rays, the light half-cone denotes reflecting rays and the two regions between them denote tunnelling rays. The left-hand curved surface denotes the core–cladding interface.

Tunnelling rays are partially reflected from the interface such that the angles of incidence and reflection are equal and the transmitted part reappears at a finite distance into the cladding (see chapter 7 of Ref. [4]). These rays are the analogue of leaky modes that have an evanescent field between the

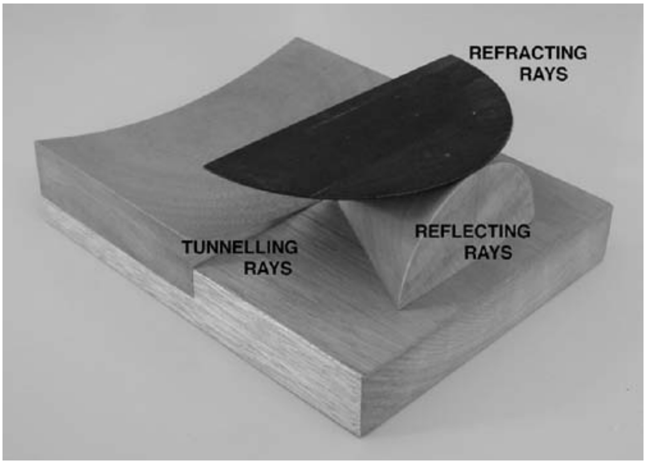


Figure A1.5.24. Representation of the three types of ray incident at a point on the curved core–cladding interface of a step-profile fibre.

core–cladding interface and the position at which the field-radiating field becomes oscillatory (see section 36.11 of Ref. [4]).

Step-profile fibre

The zig-zag ray paths of the slab waveguide become the *meridional rays* of the circular fibre and lie in planes that intersect the axis of the fibre. There is a second class of rays that follows helical-like zig-zag paths around the fibre axis without intersecting it and are know as *skew rays*. A characteristic of skew rays on the step-profile fibre is that there is an inner cylindrical surface, known as the *inner caustic*, that every straight-line segment of the ray touches periodically along its path. The radius of this caustic varies with the skewness of the ray path. The characteristic dimensions of all these ray paths can be readily quantified using simple geometry (see chapter 2 of Ref. [4]).

Graded-profile fibre

The sinusoidal-like ray paths of the graded-profile slab waveguide become the *meridional rays* of the graded-index fibre, lie in planes that intersect the axis of the fibre and are bounded by the *ray caustic*. All other rays follow continuous skew paths that have a helical-like shape, each ray path being bounded between the *outer* and *inner ray caustics* (see chapter 2 of Ref. [4]). The characteristics of each path are determined from the three-dimensional eikonal equation in cylindrical polar coordinates.

Unlike the hyperbolic-profile slab waveguide, there is no known refractive index profile that produces exactly zero dispersion in a multimode fibre for all bound ray directions. An approximately parabolic index profile fibre comes close to satisfying this goal and is known as a graded-index or GRIN fibre.

A1.5.5.6 Numerical aperture

Ray tracing provides a physical explanation for the numerical aperture (NA) of a fibre as defined by equation (A1.5.16). Using ray tracing, consider a ray from a source incident in air on the end face of a step-profile fibre at angle θ_α, as shown in figure A1.5.25. Snell's laws give

$$\sin(\theta_\alpha) = n_{\mathrm{co}} \sin(\theta_z) \tag{A1.5.68}$$

for the angle θ_z the transmitted ray makes with the fibre z-axis. Recalling that the largest angle a bound ray can subtend relative to the axis is the complementary critical angle, then the maximum angle of

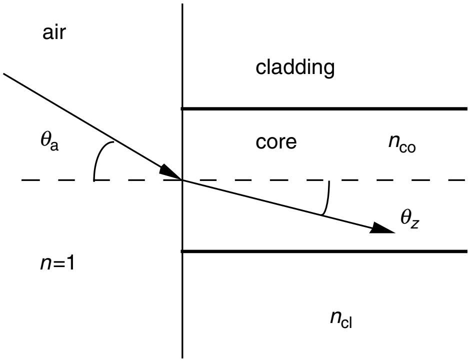


Figure A1.5.25. Ray paths refracting from air into the core of a fibre.

incidence in air for confinement within the core, θ_{max}, is given by equation (A1.5.68) with $\theta_z = \theta_c$. Recalling the definition of NA from equation (A1.5.16), this gives

$$\sin(\theta_{\text{max}}) = n_{\text{co}} \sin(\theta_c) = \text{NA} \tag{A1.5.69}$$

Assuming that NA is small, then $\theta_{\text{max}} \sim$ NA radians.

When applied to multimode fibres, this result shows that the maximum power input into bound rays (or modes) occurs when the NA is filled in a cone of angles in the range $0 < \theta_\alpha < \theta_{\text{max}}$ at every position on the end face of the core. For single-mode fibres, a ray analysis is not accurate but, nevertheless, this result suggests that filling the NA of the fibre over the core cross-section should help maximize bound mode power.

A1.5.6 Pulse dispersion

A1.5.6.1 Pulse propagation

Encoded binary information is transmitted along fibres as a sequence of discrete light pulses, normally generated by a semiconductor laser and received by a semiconductor detector at the far end of the system. Each pulse is generated by modulating the light output from the laser source, or by modulating the input into the laser itself. The bandwidth of the fibre corresponds to the number of pulses, or bits, per second, e.g. a 10 Gigabit s^{-1} bandwidth corresponds to 10^9 bits of information or pulses per second.

Dispersion is the phenomenon whereby each pulse changes its shape and effective length due to the physical properties of both the source and the fibre. As each pulse propagates along the fibre, the pulse power spreads out due to the dispersion of the fibre and its constituent materials and is also attenuated because of the losses from the fibre materials due to absorption and scattering. At the low light powers considered here (of the order of milliwatts), non-linear material effects can be neglected, i.e. dispersion caused by the effect of intense light on the fibre material, which changes the refractive index value.

The maximum distance a train of pulses can propagate data is limited because (a) the pulse spread may result in an overlap with the preceding or following pulses, so that a pulse is not unambiguously detected at the end of the fibre, leading to an error and (b) attenuation due to fibre loss is so large that a pulse has insufficient power to be detected at the end of the fibre.

The material presented here provides a background to Part C1 of this book.

Source wavelength variation

Dispersion arises because the laser source is not purely monochromatic or single wavelength; it has a finite, but small, spectral width (the variation of output power intensity with wavelength) as measured by the full width at half maximum (FWHM; figure A1.5.26). A typical laser has a spectral width that is less than one nanometre (10^{-9} metres).

To provide a qualitative description of the effect of dispersion, consider a square pulse excited by the source of excitation that has just two monochromatic wavelengths λ_1 and λ_2 (figure A1.5.27). At the beginning of the fibre, the pulses coincide. Light at the two wavelengths travels at slightly different speeds (group velocity) along the fibre, such that the pulse components at wavelengths λ_1 and λ_2 also have different speeds and thus the pulse spreads out with distance (or time). The total power in the two pulses is conserved in the absence of scattering or absorption by the fibre and the initial pulse shape becomes elongated and flattened.

To quantify the effects of dispersion on a pulse, the modal phase velocity, group velocity and group velocity dispersion need to be defined and quantified.

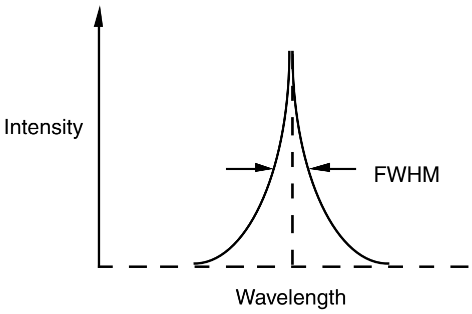


Figure A1.5.26. Schematic of laser output showing intensity variation with wavelength.

A1.5.6.2 Modal phase and group velocities

Modal phase velocity

Recall that the complete spatial and temporal dependence of the scalar fundamental mode field in the weak guidance approximation for a monochromatic source has the form

$$E(x, y, z, t) = \psi(x, y)\mathrm{e}^{\mathrm{i}(\beta z - \omega t)} \tag{A1.5.70}$$

where ω is the source frequency, related to the source (free-space) wavelength λ by $\omega = 2\pi c/\lambda$ and c is the speed of light in free space (*in vacuo*). The modal *phase velocity*, ν_{ph} is the speed with which the planar phase front of a mode in the fibre cross-section propagates along the waveguide and is expressed as

$$\nu_{\mathrm{ph}} = \frac{\omega}{\beta} = \frac{2\pi c}{\lambda\beta} = \frac{c}{n_{\mathrm{eff}}} \tag{A1.5.71}$$

where n_{eff} is the effective refractive index. The modal phase velocity is equal to the phase velocity of an infinite plane wave propagating in a uniform medium of index n_{eff}. Since the modal propagation constant β, or, equivalently, the effective index n_{eff}, depends on the source wavelength λ (or colour) through the eigenvalue equation for the fibre, the phase velocity is also wavelength dependent, i.e. $\nu_{\mathrm{ph}} = \nu_{\mathrm{ph}}(\lambda)$.

Group velocity

The speed with which the power in a pulse propagates along the fibre is given by the *group velocity*, ν_{g}, and *not* by the phase velocity. Modal group velocity is defined by the differential

$$\nu_{\mathrm{g}} = \frac{\mathrm{d}\omega}{\mathrm{d}\beta} \tag{A1.5.72}$$

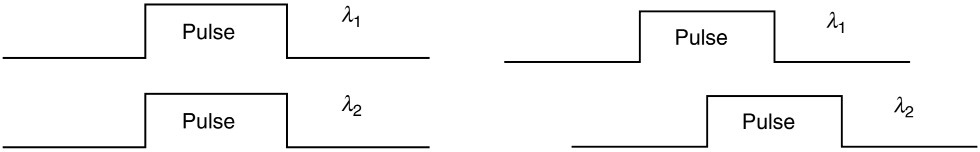


Figure A1.5.27. Pulse spread due to source spectral width.

In terms of the source wavelength λ, this definition is equivalent to

$$\nu_g = \frac{d\omega}{d\beta} = \frac{d\lambda}{d\beta}\frac{d\omega}{d\lambda} = \frac{d\lambda}{d\beta}\frac{d}{d\lambda}\left(\frac{2\pi c}{\lambda}\right) = -\frac{2\pi c}{\lambda^2}\frac{d\lambda}{d\beta} \tag{A1.5.73}$$

Setting $\beta = kn_{\text{eff}} = 2\pi n_{\text{eff}}/\lambda$ and differentiating this expression with respect to wavelength

$$\frac{d\beta}{d\lambda} = \frac{2\pi}{\lambda}\frac{dn_{\text{eff}}}{d\lambda} - \frac{2\pi n_{\text{eff}}}{\lambda^2} \tag{A1.5.74}$$

Substitution into equation (A1.5.73) leads to the expression for the modal group velocity in terms of effective index

$$\nu_g = \frac{c}{n_{\text{eff}} - \lambda\dfrac{dn_{\text{eff}}}{d\lambda}} \tag{A1.5.75}$$

The second term in the denominator accounts for the waveguide dispersion due to the finite spectral width of the source. If this term were not present then the group and phase velocities would be equal, which is the case for an infinite plane wave propagating in a uniform medium of index n_{eff}. Since β, or n_{eff}, depends on the source wavelength λ through the waveguide eigenvalue equation, the group velocity is wavelength-dependent, i.e. $\nu_g = \nu_g(\lambda)$. As the group velocity depends on the solution of the eigenvalue equation, its value is determined numerically for a general fibre profile.

A1.5.6.3 Transit time and pulse spread

If the group velocity at the central wavelength λ_c of the source of excitation is denoted by $\nu_g(\lambda_c)$, then the *transit time t* of the centre of the pulse over length L of fibre is given by

$$t = \frac{L}{\nu_g(\lambda_c)} \tag{A1.5.76}$$

In terms of the effective index n_{eff}, it is straightforward to show from equation (A1.5.75) that this expression is equivalent to

$$t = \frac{Ln_{\text{eff}}}{c} - \frac{L}{c}\lambda\frac{dn_{\text{eff}}}{d\lambda} \tag{A1.5.77}$$

where the first term on the right is the transit time for a pulse propagating the distance L in a medium of uniform index n_{eff}, and the second term accounts for the dispersion caused by the waveguide.

Group velocity dispersion

Now consider a light pulse with an initial time width τ that is excited at the beginning of the fibre. The temporal spread $\delta\tau$ in t after propagating distance L along the waveguide is due to the source spectral width $\delta\omega_s$ (in frequency units) and the corresponding spread $\delta\nu_g$ in the range of values about the mean value of the group velocity. This spread is sometimes referred to as *group velocity dispersion*. The expression for $\delta\tau$ is obtained by differentiating the transit time expression for t given above. Thus

$$\delta\tau = \frac{dt}{d\omega}\delta\omega = -\frac{L}{\nu_g^2}\frac{d\nu_g}{d\omega}\delta\omega_s = L\frac{d^2\beta}{d\omega^2}\delta\omega_s \tag{A1.5.78}$$

where the final expression on the right follows from the definition of the modal group velocity given in equation (A1.5.72).

In terms of the equivalent spectral width $\delta\lambda_s$ (in wavelength units) of the source, the above expression can be recast as

$$\delta\tau = \frac{dt}{d\lambda}\delta\lambda = -\frac{L}{\nu_g^2}\frac{d\nu_g}{d\lambda}\delta\lambda_s \tag{A1.5.79}$$

If ν_g is expressed in terms of the effective index n_{eff} of the mode using equation (A1.5.75), it is straightforward to show that

$$\delta\tau = -\frac{L\lambda}{c}\frac{d^2 n_{\mathrm{eff}}}{d\lambda^2}\delta\lambda \tag{A1.5.80}$$

If there is no net dispersion, i.e. n_{eff} is either independent of λ or varies linearly with λ, then there is no pulse dispersion.

A1.5.6.4 Sources of dispersion

When an optical pulse propagates along a practical fibre with an arbitrary index profile and cross-sectional geometry, the *total pulse dispersion* can be a combination of several different contributing effects (figure A1.5.28).

Inter-modal dispersion

This form of dispersion occurs when two or more different modes propagate simultaneously and arises because each mode has a different group velocity for the same wavelength. This phenomenon occurs regardless of the spectral width of the source. Inter-modal dispersion is the major contribution to dispersion in multimode fibres, but can be minimized using multimode fibres with a graded-index fibre as discussed in Section A1.5.5.5.

Material dispersion

In any optical material, the refractive index of the bulk material varies slightly with wavelength, i.e. $n = n(\lambda)$, in addition to any spatial variation. For example, the refractive index of the pure silica

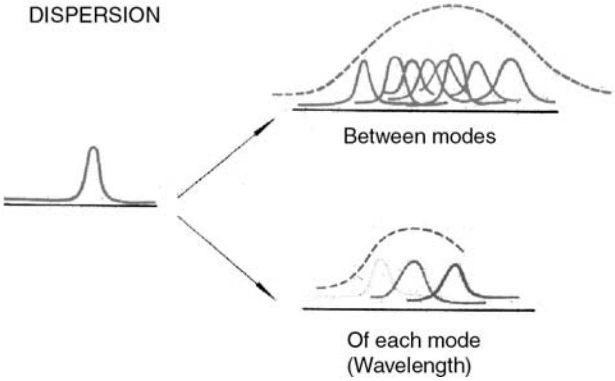


Figure A1.5.28. Dispersion includes inter-modal dispersion between different modes and the dispersion of individual modes.

cladding of a fibre decreases from 1.458 at 633 nm (wavelength of a HeNe laser) to 1.447 at 1300 nm and to 1.444 at 1550 nm. The same relative decrease in index is normally assumed for the lightly doped core material of the fibre.

Waveguide or fibre dispersion

This form of dispersion occurs in each and every mode of a fibre because of the wavelength dependence of the modal group velocity and the slight spectral width of the source. The combined effects of material and waveguide dispersion are sometimes termed *chromatic dispersion*.

Polarization-mode dispersion

In a non-circular single-mode fibre, such as a weakly guiding elliptical fibre, the two polarizations of the fundamental mode are parallel to the major and minor axes and generally have distinctly different propagation constants. Dispersion is then dominated by the inter-modal dispersion between these two polarizations. In a perfectly circular single-mode fibre, the fundamental-mode propagation constant is independent of its polarization state. Further the polarization of the mode does not change as it propagates along the fibre as long as it remains translationally invariant along its entire length.

Practical telecommunications-grade single-mode fibres are neither perfectly circular nor perfectly concentric, i.e. there are very slight non-uniformities that occur along their entire length. These non-uniformities have the effect of introducing random coupling between the two polarization states and, as a result, the accumulated group delay in each state can differ and therefore cause an overall dispersion known as *polarization-mode dispersion*. The delay difference cannot be predicted theoretically and has to be measured. In practice it is offset actively. This type of dispersion is only significant in very long single-mode fibre transmission systems over thousnds of kilometres, i.e. to submarine systems, as terrestrial systems are generally, at most, a few hundreds of kilometres long.

A1.5.6.5 Zero-dispersion, dispersion-shifted, -flattened and -compensating fibres

Zero-dispersion fibre

In principle, inter-modal dispersion can be avoided altogether if only single-mode fibres are employed. This arrangement only became a practical proposition with the development of techniques for producing low-loss single-mode fibres in the 1970s. It was then possible to design a single-mode fibre so that the combination of the fibre dispersion together with the material dispersion is zero at just one wavelength. This is a complex design procedure, as fibre and material dispersion interact with one another, i.e. they cannot be calculated separately and then simply added together. The first such designs provided a *zero-dispersion* fibre at the first fibre transmission loss minimum of 1300 nm, as shown in the left plot of figure A1.5.29.

Dispersion-shifted fibres

With the introduction of optical transmission systems operating close to the absolute minimum of silica fibre transmission loss around 1550 nm, a single-mode fibre with zero dispersion at this wavelength was required. As material dispersion is essentially fixed, an approximately triangular-shaped profile was developed to meet this need, as shown in the right plot of figure A1.5.29. Such fibres are known as *dispersion-shifted* fibres. By judicious profile design it is also possible to produce fibres that are wavelength flattened, i.e. have close to zero dispersion over the entire wavelength range of 1300–1550 nm.

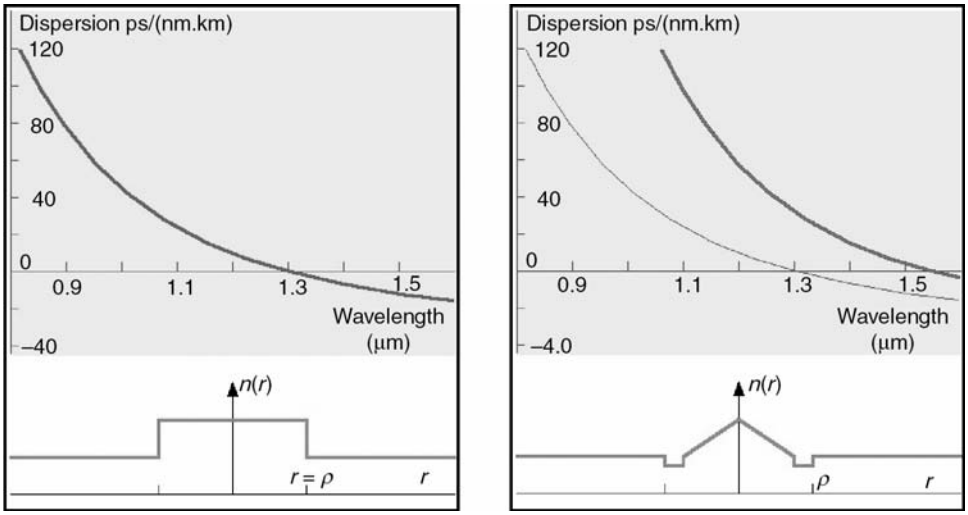


Figure A1.5.29. Plots of total dispersion against wavelength for standard fibre (left) and dispersion-shifted fibre (right).

Dispersion-compensating fibres

At the opposite extreme by combining the effects of fibre and material dispersion in such a way to enhance dispersion, single-mode fibres can be produced that have a very high and negative dispersion, of the order of $-200\,\text{ps}\,\text{nm}^{-1}\,\text{km}^{-1}$. Such fibres have application, for example, in 1300 nm transmission systems that are operated at 1550 nm to take advantage of erbium-doped optical amplifiers (EDFA). The fibres in these systems have zero dispersion at 1300 nm but a dispersion of about $20\,\text{ps}\,\text{nm}^{-1}\,\text{km}^{-1}$ when operated at 1550 nm. By inserting lengths of the large negative dispersion fibre equal to 10% of the length of the 1300 nm fibre between repeaters, the introduced dispersion can be completely offset.

A1.5.6.6 Dispersion in multimode waveguides and fibres

By definition, multimode waveguides and fibres support a large number of bound modes with propagation constants that take on almost a continuum of values in the range $kn_{\text{cl}} < \beta < kn_{\text{co}}$, where n_{cl} is the uniform cladding index and n_{co} is the maximum core index. Accordingly, pulse dispersion is strongly influenced by inter-modal dispersion but is insensitive to source wavelength. There was a flurry of research into minimizing dispersion in multimode fibres in the early 1970s prior to the development of techniques for fabricating low-loss single-mode fibres.

Step-profile multimode slab waveguide

A simple analytical example of inter-modal dispersion is provided by the step-profile slab waveguide. Consider excitation of the waveguide shown in figure A1.5.30. Assuming an incoherent source, such as a light emitting diode (LED), all bound modes are excited approximately equally, which is well modelled by assuming that all bound ray directions are equally excited within the core. As shown in Section A1.5.5.3, the range of bound ray directions for the step profile satisfies $0 < \beta_z < \beta_c$, where β_z is the angle with the z-axis.

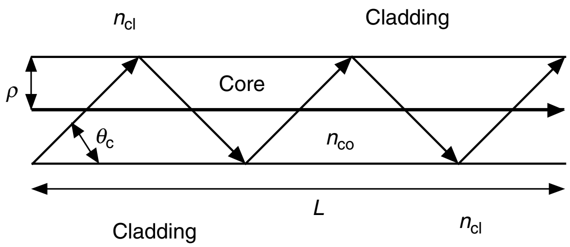


Figure A1.5.30. Bound ray path on a step-index slab waveguide.

As light propagates as a local plane wave, its phase and group velocities in the core are equal, i.e. $\nu_{ph} = \nu_g = c/n_{co}$. Accordingly, the accumulated dispersion, δt, over length L of the waveguide is given by the difference in transit times of the slowest ($\theta_z = \theta_c$) and fastest ($\theta_z = 0$) rays, which leads to

$$\delta t = \frac{Ln_{co}^2}{cn_{cl}} - \frac{Ln_{co}}{c} = \frac{L}{c}\frac{\mathrm{NA}^2}{n_{cl}} \tag{A1.5.81}$$

where NA is the numerical aperture. Note that this expression is independent of the core width. Thus, for example, for a waveguide with an NA $= 0.1$ and $n_{cl} = 1.45$, the dispersion is $23\,\mathrm{ns\,km^{-1}}$.

Zero inter-modal dispersion multimode slab waveguide

Inter-modal dispersion in slab waveguides can be totally suppressed through a judicious choice of the refractive index profile, as discussed in Section A1.5.5.4. There is a unique profile, the *hyperbolic secant* profile of equation (A1.5.67) for which the transit time for all off-axis ray paths is equal to the on-axis transit time. Note that if inter-modal dispersion is totally suppressed, both waveguide and material dispersion are still present.

Minimizing inter-modal dispersion in multimode fibres

Using ray tracing, inter-modal dispersion in circular fibres can be readily determined, by analogy, with the above analysis for the slab waveguide. Like the step-profile slab waveguide, inter-modal dispersion in a multimode step-profile fibre is relatively large. However, unlike the hyperbolic secant slab waveguide, there is no known profile for the fibre that leads to exactly zero inter-modal dispersion when all ray directions, both meridional and skew, are excited.

Nevertheless, it is possible to minimize inter-modal dispersion using a core profile that is approximately parabolic, as might be anticipated from the approximately parabolic variation of the hyperbolic secant profile. The term *graded-index fibre* is often used to denote a multimode fibre with this core profile surrounded by the normal uniform cladding. However, the bandwidth of a single-mode fibre is still significantly larger than that of a graded-index fibre.

A1.5.7 Bend loss

A mode propagating on a straight fibre or waveguide fabricated from non-absorbing, non-scattering materials will in principle propagate indefinitely without any loss of power. However, if a bend is introduced, the translational invariance is broken and power is lost from the mode as it propagates into, along and out of the bend. This applies to the fundamental mode in the case of single-mode fibres and waveguides and to all bound modes in the case of bent multimode fibres or waveguides.

It is conventional to distinguish between the two types of losses that can occur on a bend. *Transition loss* is associated with the abrupt or rapid change in curvature at the beginning and the end of a bend, and *pure bend loss* is associated with the loss from the bend of constant curvature in between.

A1.5.7.1 Transition loss

Consider an abrupt change in the curvature κ from the straight waveguide ($\kappa = 0$) to that of the bent waveguide of constant radius R_b ($\kappa = 1/R_\mathrm{b}$). The fundamental-mode field is shifted slightly outwards in the plane of the bend, thereby causing a miss-match with the field of the straight waveguide. The fractional loss in fundamental-mode power, $\delta P/P$, can be calculated from the overlap integral between the fields. Within the Gaussian approximation to the fundamental mode field and assuming that the spot size s and core radius or half-width ρ are approximately equal, this gives

$$\frac{\delta P}{P} \approx \frac{1}{16}\frac{V^4}{\Delta^2}\frac{\rho^2}{R_\mathrm{b}^2} \tag{A1.5.82}$$

where V is the fibre or waveguide parameter and Δ is the relative index difference. For typical values of $V = 2.5$, $\Delta = 0.005$, $\rho = 5\,\mathrm{mm}$ and a bend radius of 5 mm, the transition loss is about 9.8% of fundamental-mode power.

Minimizing transition loss

There are a couple of strategies for significantly reducing transition loss. In the case of planar waveguides it is often possible to fabricate the bend so that there is an abrupt offset between the cores of the straight and bent waveguides in the plane of the bend. In figure A1.5.31 this can be seen as being equivalent to displacing the bent core downwards so that the two fundamental-mode fields overlap. Alternatively, if a gradual increase in curvature is introduced between the straight and uniformly bent sections, the fundamental field of the straight waveguide will evolve approximately adiabatically into the offset field of the uniformly bent section.

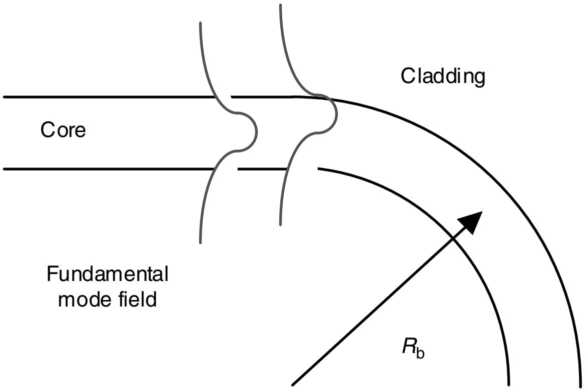


Figure A1.5.31. Outward shift in the fundamental-mode electric field on entering a bend.

A1.5.7.2 Pure bend loss

In following the curved path of the core of constant radius R_b, the fundamental mode continuously loses power to radiation, known as *pure bend loss*. In assuming that the cladding is essentially unbounded, the physical effect is analogous to radiation from a bent antenna in free space. The radiation increases rapidly with decreasing bend radius and occurs predominantly in the plane of the bend; in any other plane the effective bend radius is larger and hence the loss is very much reduced (figure A1.5.32).

The physical basis of bend loss can be interpreted in terms of the motion of the planar phase front of the fundamental mode as it rotates about an axis through the centre of the bend at C with angular velocity Ω. On the straight waveguide, the phase front travels with phase velocity $v_\phi = \omega/\beta$, where ω is the source frequency and β the propagation constant. By matching the phase velocities of the straight and bent waveguides on the fibre axis at the beginning of the bend

$$R_b\Omega = \frac{\omega}{\beta} \Rightarrow \Omega = \frac{\omega}{\beta R_b} = \frac{2\pi c}{\lambda\beta R_b} = \frac{c}{n_{\mathrm{eff}}R_b} \qquad \text{(A1.5.83)}$$

in terms of the source wavelength λ and the fundamental mode effective index n_{eff}. Along the planar phase front through C, the phase velocity increases monotonically with the distance from C, until at a radius R_{rad}, the phase velocity is equal to the speed of light in the cladding, i.e. when

$$\Omega R_{\mathrm{rad}} = \frac{c}{n_{\mathrm{cl}}} \Rightarrow R_{\mathrm{rad}} = \frac{c}{\Omega n_{\mathrm{cl}}} = \frac{n_{\mathrm{eff}}}{n_{\mathrm{cl}}}R_b \qquad \text{(A1.5.84)}$$

Since by definition $n_{\mathrm{eff}} > n_{\mathrm{cl}}$, it follows that $R_{\mathrm{rad}} > R_b$.

Radiation caustic

The phase velocity anywhere on the modal phase front rotating around the bend cannot exceed the speed of light in the cladding. Hence, beyond radius R_{rad} the modal field must necessarily radiate into the cladding, the radiation being emitted tangentially. The interface between the guided portion of the modal field around the bend and the radiated portion at R_{rad} is known as the *radiation caustic*, and is the apparent origin of radiation.

Between the core and the radiation caustic, the modal field is evanescent and decreases approximately exponentially with increasing radial distance from C. As the bend radius increases, the radiation caustic moves farther into the cladding, and the level of radiated power decreases. Conversely,

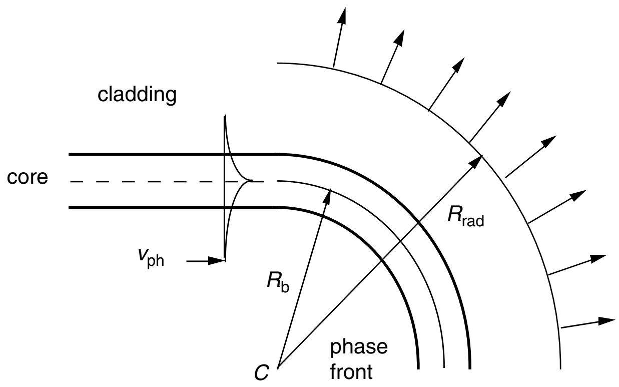


Figure A1.5.32. Rotation of the mode phase front along a bend.

as the bend radius decreases, the radiation caustic moves closer to the core and the radiation loss increases. In practical fibres, the radiated power from the bend is either absorbed by the acrylic coating surrounding the outside of the cladding or propagates through the coating into free space.

Mode attenuation

If z is the distance from the beginning of the bend relative to the fibre axis, then the total fundamental-mode power $P(z)$ attenuates according to

$$P(z) = P(0)\mathrm{e}^{-\gamma z} \tag{A1.5.85}$$

where $P(0)$ is the total fundamental mode power entering the bend and γ is the power attenuation coefficient. In decibels, this relationship is equivalent to

$$\mathrm{dB} = 10\log_{10}\left(\frac{P(z)}{P(0)}\right) = 10\log_{10}(\mathrm{e}^{-\gamma z}) = (20\log_{10}\mathrm{e})\gamma z = 8.686\gamma z \tag{A1.5.86}$$

indicating that the loss of power per unit length of bent fibre is $8.686\gamma\,\mathrm{dB}$.

Step-profile fibre

In terms of the core and cladding modal parameters U and W, respectively, relative index difference Δ, core radius ρ, fibre parameter V and the bend radius R_b, an approximate expression for γ for the fundamental mode of a step-profile fibre has the form (see chapter 23 of Ref. [4])

$$\gamma = \left(\frac{\pi\rho}{R_\mathrm{b}}\right)^{1/2}\frac{V^2W^{1/2}}{2\rho U^2}\exp\left\{-\frac{4}{3}\Delta\frac{R_\mathrm{b}}{\rho}\frac{W^3}{V^2}\right\} \tag{A1.5.87}$$

where R_b is necessarily large compared to ρ because it is not possible to bend a fibre into a radius much below 1 cm without breakage. The pure bend loss coefficient is most sensitive to the expression inside the exponent because $R_\mathrm{b} \gg \rho$. Loss decreases very rapidly with increasing values of R_b or Δ or V (since W also increases with V), and becomes arbitrarily small as $R_\mathrm{b} \to \infty$.

Consider, for example, a standard single-mode fibre such as the Corning SMF28, for which $\Delta = 0.3\%$, $\rho = 4\,\mu\mathrm{m}$, and $\mathrm{NA} = 0.12$. At a wavelength of $1.3\,\mu\mathrm{m}$, the fibre parameter is $V = 2.32$. The eigenvalue equation for the step profile gives $U = 1.66$, $W = 1.62$ and the power attenuation coefficient of equation (A1.5.87) reduces to

$$\gamma = \frac{3.48\times 10^4}{\sqrt{R_\mathrm{b}}}\exp(-0.78R_\mathrm{b}) \tag{A1.5.88}$$

where R_b is in millimetres and γ has units of m^{-1}. Accordingly, the fractional power loss from the fundamental mode due to pure bend loss in a single loop of fibre is expressible as

$$\frac{P(0) - P(2\pi R_\mathrm{b})}{P(0)} = 1 - \exp\left(-\frac{\pi\gamma R_\mathrm{b}}{500}\right) \tag{A1.5.89}$$

where R_b is in millimetres. Thus, if $R_\mathrm{b} = 150\,\mathrm{mm}$ (the radius of the standard drum on which fibre is normally spooled), then $\gamma \sim 10^{47}\,\mathrm{m}^{-1}$ and loss is totally negligible, but if $R_\mathrm{b} = 10\,\mathrm{mm}$, then $\gamma = 4.5\,\mathrm{m}^{-1}$ and the loss is almost 25% in one loop of fibre.

A1.5.7.3 Bend loss in multimode fibres and waveguides

Uniformly bent multimode waveguides lose power because each mode is attenuated as it propagates around the bend. Like the fundamental mode, the power attenuation coefficient for each mode is dominated by the common exponential dependence

$$\exp\left\{-\frac{4}{3}\Delta\frac{R_{\mathrm{b}}}{\rho}\frac{W^3}{V^2}\right\} = \exp\left\{-\frac{4}{3}kR_{\mathrm{b}}\Delta\frac{(n_{\mathrm{eff}}^2 - n_{\mathrm{cl}}^2)^{3/2}}{(n_{\mathrm{co}}^2 - n_{\mathrm{cl}}^2)}\right\} \tag{A1.5.90}$$

where the right-hand expression follows by setting $\beta = kn_{\mathrm{eff}}$ and $V = k\rho(n_{\mathrm{co}}^2 - n_{\mathrm{cl}}^2)^{1/2}$.

Recall that the effective index of every bound mode of the straight fibre mode lies in the range $n_{\mathrm{cl}} < n_{\mathrm{eff}} < n_{\mathrm{co}}$, the lower-order modes having n_{eff} values closer to n_{co} and the higher-order modes having n_{eff} values closer to n_{cl}. It follows from equation (A1.5.90) that the higher-order modes are attenuated much more quickly than the lower-order modes. However, the overall attenuation along the bent multimode waveguide is determined by the attenuation of each mode. Since the equilibrium distribution of power between modes also depends on the slight variations along the waveguide, the overall attenuation around a bend must take account of this distribution. As a result, there is no simple analytical expression for bend loss in multimode waveguides.

Numerical quantification of bend loss for multimode slab waveguides and fibres with step and parabolic profiles have been undertaken using a combination of ray tracing and loss coefficients (figure A1.5.33). The loss coefficients account for bend loss at reflections from the core–cladding interface, in the case of the step profile, and from the outer caustic in the case of the parabolic profile. For the slab waveguide this involves a two-dimensional integration over the core width and the range of ray directions, while the fibre involves a four-dimensional integration of the core cross-sectional area and all meridional and skew ray paths (see chapter 9 of Ref. [4]).

A1.5.8 Excitation, reflection, mismatch, offset and tilt

Waveguides and fibres are illuminated by a variety of sources, including lasers, light emitting diodes or light propagating out of another fibre or waveguide. Because of the relatively small core size of fibres and

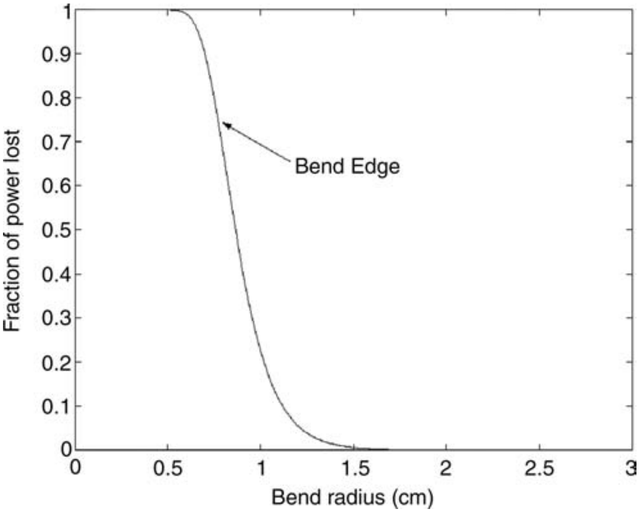


Figure A1.5.33. Bend loss as a function of bend radius.

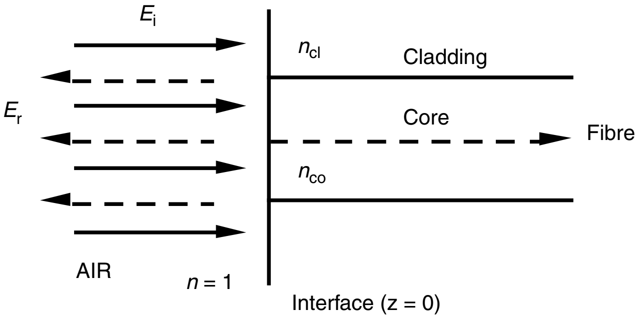


Figure A1.5.34. Reflection of light from the end of a fibre or waveguide.

waveguides, it is necessary to use an intense source to ensure that sufficient light enters the fibre to be detectable at the far end. The goal is normally to maximize the fraction of incident light that can excite the fundamental mode in the case of a single-mode fibre or all the bound modes in the case of a multimode fibre.

However, an analysis of excitation is a complex issue because of (a) reflection from the fibre end face, (b) mismatch between the source and fibre modes, (c) offset between the source and the fibre, and (d) tilt between the source and the fibre. Further, an exact mathematical formulation of the problem, even within the weak guidance approximation, would require an appropriate representation of both the reflected and transmitted bound mode and radiation fields, and the solution of an extremely complex and large matrix of equations. Fortunately for most practical applications, it is not necessary to undertake such an analysis. Here we provide some insight and methodology that provides an accurate quantification of all these effects for most practical applications.

A1.5.8.1 End-face reflection

If a fibre is excited by light propagating in air, the light will be partially reflected because of the difference in indices between air and the fibre material. Consider the simple situation shown in figure A1.5.34 where a plane wave, representing the incident light, impinges normally on the end face of the fibre with a scalar transverse electric field E_i. A fraction of the field is reflected into the backward-propagating field E_r and the balance is transmitted into the fibre or waveguide.

A simple way to estimate the fraction of reflected light is to approximate the fibre or waveguide core–cladding cross-section by a uniform, infinite medium of refractive index approximately equal to the core index n_{co} since the core and cladding indices are similar in the case of weak guidance. For the incident plane wave, Fresnel's laws determine the fraction of reflected light power according to

$$\left(\frac{n_{co} - 1}{n_{co} + 1}\right)^2 \qquad \text{(A1.5.91)}$$

For a typical fibre core with index $n_{co} = 1.46$, this formula gives around 3.5% for the fraction of incident power reflected. It should also be noted that the same fraction of light will be reflected from the propagating light impinging on the end face at the far end of the fibre, assuming it is also located in air.

A1.5.8.2 Anti-reflecting coating

The reflection of light from the end face of a fibre can be suppressed by depositing a thin anti-reflection coating with index different from that of the fibre, as shown schematically in figure A1.5.35. The coating

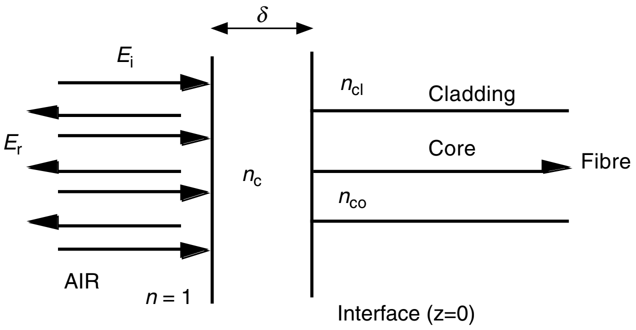


Figure A1.5.35. Schematic of an anti-reflecting coating on the end of a fibre or waveguide.

has uniform thickness δ and refractive index n_c. A simple plane-wave analysis of propagation across this layer at normal incidence from air into the fibre shows that reflection can be totally suppressed from the end face between the coating and air provided that the following two conditions are satisfied

$$d = \frac{m\lambda}{4n_c}, \ m = 1, 3, 5, \ldots; \quad n_c = \sqrt{n} \tag{A1.5.92}$$

where $m = 1,3,5,\ldots$, n is the mean fibre index and λ is the source wavelength.

A1.5.8.3 Mode excitation

Assume that the transmitted scalar electric field of the source on the fibre or waveguide side of the interface at $z = 0$ is given is by the scalar expression $\psi_s(x, y)$, where x, y are Cartesian coordinates in the cross-section (figure A1.5.36). This function is assumed to take into account any reflection loss from the interface and also contains the implicit monochromatic time dependence $\exp(-i\omega t)$.

The complete field everywhere within the core and cladding for $z > 0$ consists, in general, of the sum of the discrete set of bound modes, and the radiation field. Since radiated power is lost from the fibre through the cladding, the total bound-mode field, $\psi_s(x, y)$, is expressible as a sum over all bound modes

$$\psi_s(x, y) = \sum_{m=1}^{M} a_m \psi_m(x, y) e^{i\beta_m z} \tag{A1.5.93}$$

where the a_m are amplitude constants, $\psi_m(x, y)$ is the solution of the scalar wave equation for the field of the mth mode, β_m is the propagation constant for the mth mode and the summation is over all M-bound modes. For a single-mode fibre, $M = m = 1$, and the polarization of the exciting field is assumed to be parallel to the mode polarization.

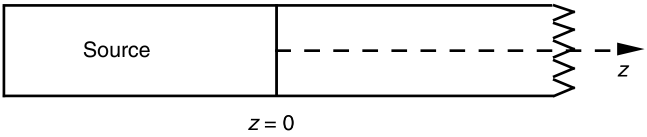


Figure A1.5.36. Schematic of a source abutting the end of a fibre or waveguide at $z = 0$.

Modal amplitudes

The expansion over bound modes and the radiation field ψ_{rad} must match the transmitted field on the end face at $z = 0$. Hence

$$\psi_{\mathrm{s}}(x, y) = \sum_{m=1}^{M} a_m \psi_m(x, y) + \psi_{\mathrm{rad}} \tag{A1.5.94}$$

The amplitude of the kth bound mode is then obtained by multiplying this equation on both sides by $\psi_k(x, y)$, and integrating over the infinite cross-section A_∞. Since, by definition, any bound-mode field is necessarily orthogonal to the radiation field

$$\int_{A_\infty} \Psi_{\mathrm{s}}(x, y)\psi_k(x, y)\,\mathrm{d}x\,\mathrm{d}y = \sum_{m=1}^{M} a_m \int_{A_\infty} \psi_m(x, y)\psi_k(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{A1.5.95}$$

The amplitude coefficient a_m of each mode is then determined by using the orthogonality property of the scalar modal fields

$$\int_{A_\infty} \psi_m(x, y)\psi_k(x, y)\,\mathrm{d}x\,\mathrm{d}y = 0 \quad \text{if}\, m \neq k \tag{A1.5.96}$$

Accordingly the right-hand side of equation (A1.5.95) reduces to a single term when $k = m$ and the amplitude of the kth excited mode is given by

$$a_k = \int_{A_\infty} \Psi_{\mathrm{s}}(x, y)\psi_k(x, y)\,\mathrm{d}x\,\mathrm{d}y \Big/ \int_{A_\infty} \psi_k(x, y)^2\,\mathrm{d}x\,\mathrm{d}y \tag{A1.5.97}$$

Note that for a circular fibre the Cartesian element of area $\mathrm{d}x\,\mathrm{d}y \rightarrow r\mathrm{d}r\mathrm{d}\phi$ in polar coordinates with $0 < r < \infty$ and $0 < \phi < 2\pi$.

Modal power

Equation (A1.5.28) determines that the total power P_k propagating in each excited mode is proportional to

$$(a_k)^2 \int_{A_\infty} (\psi_k)^2\,\mathrm{d}x\,\mathrm{d}y \tag{A1.5.98}$$

where A_∞ is the infinite cross-section of the fibre or waveguide and $k = 1, 2, \ldots, M$. On substituting a_k from equation (A1.5.97) the fraction of transmitted power in each excited mode relative to the total power exciting the fibre is given by

$$\frac{\left(\displaystyle\int_{A_\infty} \psi_k \Psi_{\mathrm{s}}\,\mathrm{d}x\,\mathrm{d}y\right)^2}{\displaystyle\int_{A_\infty} (\psi_k)^2\,\mathrm{d}x\,\mathrm{d}y \quad \int_{A_\infty} (\Psi_{\mathrm{s}})^2\,\mathrm{d}x\,\mathrm{d}y} \tag{A1.5.99}$$

The total fraction of incident power that is transmitted into bound modes is given by summing equation (A1.5.99) over all values of k. Thus, the fraction of total exciting power lost as radiation is given by

$$1-\sum_{k=1}^{M}\frac{\left(\displaystyle\int_{A_\infty}\psi_k\Psi_{\mathrm{s}}\,\mathrm{d}x\,\mathrm{d}y\right)^2}{\displaystyle\int_{A_\infty}(\psi_k)^2\,\mathrm{d}x\,\mathrm{d}y\int_{A_\infty}(\Psi_{\mathrm{s}})^2\,\mathrm{d}x\,\mathrm{d}y} \tag{A1.5.100}$$

A1.5.8.4 Fibre splicing, miss-match, offset and tilt

Optical fibres are normally spliced together by removing the coating (mechanically or chemically), cleaving the fibre ends so that their end faces are flat and perpendicular to the fibre axis and butting them together in a V-groove. The fibres are then heated by an electric arc to soften the glass and fuse them together to form a smooth continuous joint. Splicing fibres using this technique essentially requires fibres with the same cladding diameter, commonly referred to as the *outer diameter* or *o/d*, although their core refractive index profiles and radii can differ quite significantly.

Splice loss

The difference between the cross-sectional profile and geometry of the cores of the two fibres is known as the *miss-match*. For single-mode fibres the miss-match accounts for the loss of power at the splice as the fundamental mode fields of the two fibres will not be identical. Additional losses will occur if the fibres are *offset* transversely, or if their axes are no longer parallel and have a relative *tilt*.

Splice loss, due to any of the three loss mechanisms shown schematically in figure A1.5.37, can formally be calculated and quantified using the analysis of mode excitation in the previous section but with the input source being the field of the fundamental mode of the input fibre. However, such an analysis would require detailed knowledge of the refractive index profiles of both fibres and formal solution of the scalar wave equation (numerically) to determine the fields of the fundamental mode of each fibre. A simpler approach, which requires knowledge only of each fibre's spot size (taken from fibre measurement) and gives an approximate but accurate measure of splice loss, is to use the Gaussian approximation of Section A1.5.4.7 to represent the field of the fundamental mode of each fibre.

Miss-match loss

Accordingly, if ψ_1 and ψ_2 denote the fundamental fields of fibres 1 and 2, respectively, in the weak-guidance approximation then the representation of the Gaussian approximation is

$$\psi_1(r)=a_1\exp\left\{-\frac{r^2}{2s_1^2}\right\};\quad \psi_2(r)=a_2\exp\left\{-\frac{r^2}{2s_2^2}\right\} \tag{A1.5.101}$$

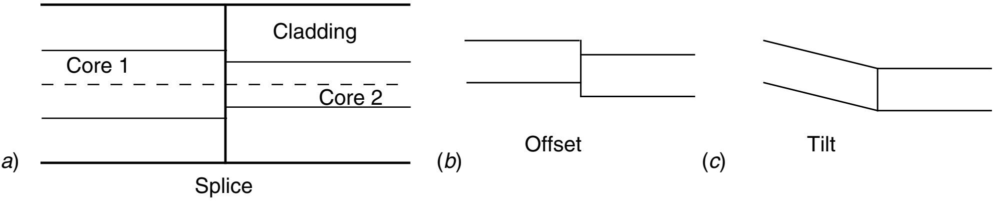


Figure A1.5.37. (*a*) Miss-match, (*b*) offset and (*c*) tilt between fibre cores.

where s_1 and s_2, a_1 and a_2 are, respectively, the spot sizes and amplitudes for fibre 1 and fibre 2, and r is the cylindrical radial coordinate measured from the fibre axis.

The power P_1 propagating in the fundamental mode in fibre 1 is obtained by integrating the local power flow power density parallel to the fibre axis over the infinite cross-section relative to polar coordinates (r, ϕ) in the fibre cross-section. In weak guidance the power flow density in the z-direction is given by equation (A1.5.28). Since ψ_1 is axisymmetric, the ϕ integration is readily performed and leads to

$$P_1 = \frac{n_{\mathrm{co}}}{2}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2}\int_0^{2\pi} r\psi_1^2 \mathrm{d}r\,\mathrm{d}\phi = \pi n_{\mathrm{co}}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2}\int_0^{\infty} r\psi_1^2 \mathrm{d}r \tag{A1.5.102}$$

Substituting for ψ_1 leads to a standard integral

$$P_1 = \pi(a_1)^2 n_{\mathrm{co}}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2}\int_0^{\infty} r\exp\left\{-\frac{r^2}{s^2}\right\}\mathrm{d}r = \frac{\pi n_{\mathrm{co}}}{2}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} s_1^2 a_1^2 \tag{A1.5.103}$$

Similarly, the power P_2 propagating in the fundamental mode in fibre 2 is given by

$$P_2 = \frac{\pi n_{\mathrm{co}}}{2}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} s_2^2 a_2^2 \tag{A1.5.104}$$

There is negligible reflected power at the splice because both fibres are weakly guiding and their refractive index profile values are very similar. In the pure silica cladding the indices are of course identical.

The amplitude of the transmitted mode, a_2, is determined by using fundamental mode field ψ_1 in fibre 1 as the source of excitation (ψ_s) of the fundamental mode ψ_2 (ψ_k) for fibre 2 in equation (A1.5.97), leading to

$$a_2 = \frac{\displaystyle\int_{A_\infty} \Psi_1\psi_2 \mathrm{d}x\,\mathrm{d}y}{\displaystyle\int_{A_\infty} \psi_2^2 \mathrm{d}x\,\mathrm{d}y} = a_1\frac{\displaystyle\int_0^{\infty} r\exp\left(-\frac{r^2}{2}\left[\frac{1}{s_1^2}+\frac{1}{s_2^2}\right]\right)\mathrm{d}r}{\displaystyle\int_0^{\infty} r\exp\left(-\frac{r^2}{s_2^2}\right)\mathrm{d}r} \tag{A1.5.105}$$

By analogy with the derivation of the power integrals above, these integrals are readily evaluated and lead to

$$a_2 = a_1\frac{2s_1^2}{s_1^2+s_2^2} \tag{A1.5.106}$$

Using equations (A1.5.105) and (A1.5.106), the fractions of transmitted power and the fraction of power lost are given, respectively, by

$$\frac{P_2}{P_1} = \frac{s_2^2 a_2^2}{s_1^2 a_1^2} = \frac{4s_1^4}{(s_1^2+s_2^2)^2}\frac{s_2^2}{s_1^2}; \quad 1-\frac{P_2}{P_1} = 1-\frac{4s_1^4}{(s_1^2+s_2^2)^2}\frac{s_2^2}{s_1^2} = \left(\frac{s_1^2-s_2^2}{s_1^2+s_2^2}\right)^2 \tag{A1.5.107}$$

Thus, for example, a 1% relative difference in the two spot sizes gives rise to a power loss of only 0.01%, whereas a 10% relative difference gives a 1% loss.

Offset and tilt losses

If two identical single-mode fibres with the same spot size s are offset parallel to one another by distance d, a similar analysis to the Gaussian approximation for the miss-match loss gives the first expression in equation (A1.5.108) for the fraction of power loss. Similarly if the same two fibres have a relative tilt angle θ_s, the second expression in equation (A1.5.108) determines the fraction of power loss

$$1 - \exp\left(-\frac{d^2}{2s^2}\right); \quad 1 - \exp\left(-\frac{(kn_{\text{co}}\theta_s)^2}{2}\right) \tag{A1.5.108}$$

where k is the free-space wavenumber and n_{co} is the core index value. If both offset and tilt losses are present, then the two expressions in equation (A1.5.108) are multiplied together to determine the overall loss. Similarly, if miss-match loss is also present, then a third multiplicative factor given by the second expression in equation (A1.5.107) should be included.

A1.5.8.5 Near and far fields

The *near field* refers to the field of a fibre or waveguide on the far end face in air. In the case of a sufficiently long length of single-mode fibre, the near field will be virtually identical in shape to the fundamental-mode field within the core and cladding because the power in all higher-order cladding modes will have been absorbed and scattered by the coating material. However, the amplitude of the near field will reduced by about 1.75% to account for the 3.5% of reflected power from the end face back along the fibre. For multimode fibres, ray tracing and Snell's laws can be used to determine the near-field pattern.

Suppression of end-face reflection

When light propagating in the excited bound mode(s) of a fibre reaches the far end of the fibre, it is partly reflected back along the fibre because of the change in index from glass to air at the end face. The 3.5% of reflected power propagates back to the front face and all but 3.5% of this light is transmitted back into the source laser where it can interfere with the operation of the laser.

This reflection problem can be avoided in practice by cleaving the end face at a sufficiently large angle, typically 10°, to deflect the reflected light into the cladding and out of the fibre core as illustrated in figure A1.5.38.

Far field

The *far field* of a fibre or waveguide is the field emerging from its far end into air as measured at a sufficiently large distance from the end face. In the case of multimode fibres, the shape of the far field can be predicted using ray tracing by tracking the rays exiting the end face and taking into account their refraction at the end face.

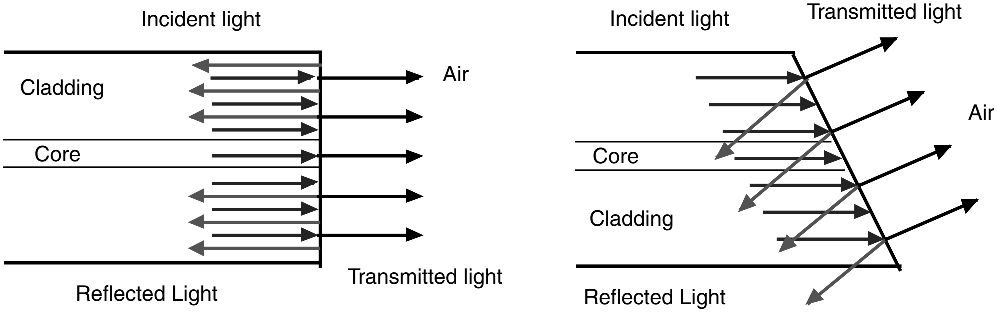


Figure A1.5.38. Reflection from the far end face of a fibre without and with an angle face.

For single-mode fibres and waveguides with relatively small core sizes of the order of a few wavelengths of the source light, the far field is *diffracted* at the end face. Because of the confining effect of the cladding on light in the core, this effect is analogous to the diffraction of a beam of light passing through a small orifice and then spreading laterally to form a series of lobes when viewed sufficiently far from the end face.

If the first lobe makes a far-field diffraction angle θ_d with the fibre axis, a shown in figure A1.5.39, standard diffraction theory can be used to relate this angle to the fibre parameters.

A1.5.9 Perturbed and non-uniform fibres and waveguides

There are two sets of problems in which (a) the cross-sectional geometry and/or the refractive index profile of a fibre or waveguide are subject to uniform longitudinal perturbations and (b) the translational invariance of a waveguide or fibre no longer holds, but a quasi-modal description of propagation is still applicable. In the first set, translational invariance still holds so that the structure supports bound modes, whereas in the second set, which includes longitudinal tapering, mode coupling necessarily occurs. In keeping with the thrust of earlier sections, an outline of the methodology employed in each of these techniques will be presented within the weak guidance approximation.

A1.5.9.1 Uniform perturbations: single core

When a fibre or waveguide has a single core for which the field and propagation constant are known for any bound mode, a simple perturbation technique enables the propagation constant to be determined when a uniform perturbation in the cross-sectional geometry and/or refractive index profile is applied to the length of the fibre or waveguide.

Consider a waveguide or fibre with known profile $n(x, y)$, scalar modal field $\psi(x, y)$ and propagation constant β, respectively. If the waveguide or fibre is slightly but uniformly distorted along its length so that it now has profile $\bar{n}(x, y)$, field $\bar{\psi}(x, y)$ and propagation constant $\bar{\beta}$, respectively, then the perturbed propagation constant can be expressed approximately in terms of the unperturbed quantities according to

$$\bar{\beta}^2 = \beta^2 + k^2 \frac{\displaystyle\int_{A_\infty} (\bar{n}^2 - n^2)\Psi^2 \mathrm{d}A}{\displaystyle\int_{A_\infty} \psi^2 \mathrm{d}A} \qquad \text{(A1.5.109)}$$

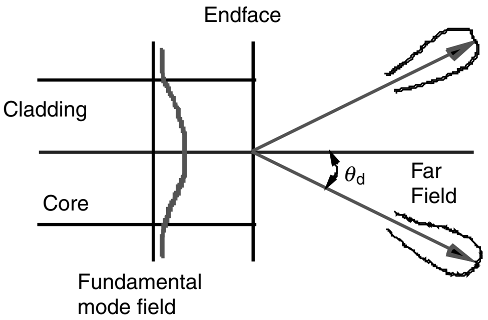


Figure A1.5.39. Fundamental mode evolving into lobes in the far field.

where k is the free-space wavenumber and the integrals are over the infinite cross-section of the waveguide or fibre (see chapter 18 of Ref. [4]).

Slight absorption

To illustrate the usefulness of the above result, consider a step-profile fibre with a loss-less cladding and core which is slightly absorbing, i.e. n_{co} is replaced by the complex expression $n_{\mathrm{co}} + \mathrm{i}\delta$, where $\delta << 1$ is a constant. By substituting into equation (A1.5.109) to the lowest order in δ gives

$$\bar{\beta}^2 = \beta^2 + 2i\delta k^2 n_{\mathrm{co}} \frac{\int_{A_{\mathrm{co}}} \psi^2 \mathrm{d}A}{\int_{A_\infty} \psi^2 \mathrm{d}A} = \beta^2 + 2i\delta\eta k^2 n_{\mathrm{co}} \tag{A1.5.110}$$

where A_{co} is the core cross-sectional area and η is the fraction of mode power propagating in the core. Upon taking the square root, setting $\beta \sim kn_{\mathrm{co}}$, then ordering δ

$$\bar{\beta} = \beta\left(1 + \frac{2i\delta\eta k^2 n_{\mathrm{co}}}{\beta^2}\right)^{1/2} \cong \beta + \frac{i\delta\eta k^2 n_{\mathrm{co}}}{\beta} \cong \beta + i\delta\eta k \tag{A1.5.111}$$

Hence the mode power $P(z)$ at distance z along the fibre is attenuated according to

$$P(z) = P(0)\exp(-2\delta\eta kz); \quad \mathrm{dB} = 8.686\delta\eta kz \tag{A1.5.112}$$

Thus, the attenuation increases with increasing absorption but decreases with increasing mode order because a larger fraction of power propagates in the cladding. In other words, the effect of absorption is most pronounced for the fundamental mode.

A1.5.9.2 Uniform perturbations: multiple cores

When more than one core is present in a fibre or waveguide, there are two essentially equivalent techniques for analysing propagation: *super-modes* and *coupled-mode theory*. This is brief background material for the more detailed applications to light processing devices discussed in Chapter B3 of this book.

Super-mode analysis

By analogy with the discussion in Section A1.5.3.10, super-modes are the bound modes associated with the multiple cores such that propagation of guided optical power can be expressed in terms of a summation over the super-modes. For example, in the case of two identical cores that in isolation from one another are single mode, there are two super-modes with even and odd symmetry, respectively, relative to the mid-plane between the two cores. If both modes are excited simultaneously, there will be interference or beating between them because of the difference in their propagation constants. Physically this leads to a coupling of optical power between the two cores.

Coupled-mode analysis

An alternative approach is to employ coupled-mode theory, the details of which are described in more detail in Section A1.5.9.4 below. Physically, this approach can be thought of as follows. When there is only one core present, the fundamental mode propagates unperturbed. If a second parallel core is now introduced supporting a second fundamental mode with a propagation constant identical to that in the first core, this situation gives rise to a resonance phenomenon and power is transferred from the first core to the second core and back again while the modes propagate. This periodic transfer of power between the two cores can be described quantitatively using coupled-mode equations and the results are formally identical to those derived using super-modes.

A1.5.9.3 Tapered fibres and waveguides

A waveguide or fibre taper corresponds to a variation of the core cross-section and/or refractive index profile with distance z along the fibre, i.e. $\rho = \rho(z)$ and/or $n = n(z)$, where ρ is either the half-side of a square-core waveguide or the the core radius of a step-profile fibre and n is the profile. If the taper is single-mode, then at each position z along the taper, only the fundamental mode is bound, but because of the variation in $\rho(z)$ the fundamental mode must lose power by coupling to radiation. However it is intuitive that if the core radius $\rho(z)$ varies sufficiently slowly along the taper, than the radiation loss should be minimal. If there were no power loss, then the taper would be described as *adiabatic*; with a small but finite loss, the taper can be described as *approximately adiabatic*.

A1.5.9.4 Local modes

If the taper is approximately adiabatic, and the fibre or the waveguide is single-mode along its entire length, it is intuitive that the fundamental mode defined by the core cross-section at the beginning of the taper will have virtually the same power as the fundamental mode at the end of the taper. However, its scalar field $\psi(x, y)$ and propagation constant β at the ends of the taper can be quite different, so how can we accommodate this change and at the same time retain the fundamental mode description?

The answer is the concept of a *local mode*. Such a mode has the property that, at each position along the taper, its modal field $\psi(x, y, z)$ and propagation constant $\beta(z)$ adapt so that they are determined by the local cross-sectional geometry and index profile. Hence at cross-section AA′ in figure A1.5.40 the core radius is, say, $\rho(z_A)$, while at BB′ it will be $\rho(z_B)$. The solutions of the scalar wave equation for the field and propagation constant are then determined for the step profile with radius $\rho(z_A)$ at AA′ and radius $\rho(z_B)$ at BB′. To connect these solutions at each end, consider the local mode phase and amplitude variation separately.

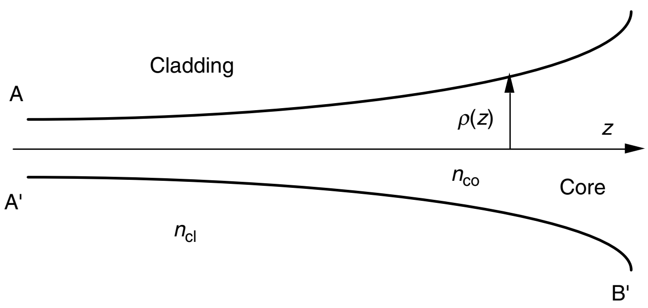


Figure A1.5.40. Schematic of a tapered core.

Local mode phase and power dependence

In a uniform fibre, phase increases linearly with distance as βz. If β now varies continuously with z, i.e. $\beta \rightarrow \beta(z)$, it is intuitive that the linear accumulation of phase is replaced by a continuous summation that can be represented by the integral

$$\int_z \beta(z)\,\mathrm{d}z \tag{A1.5.113}$$

that reduces to βz when $\beta(z)$ is a constant.

In weak guidance the total power P propagating in the fundamental mode along a uniform waveguide or fibre is given in terms of its amplitude a_0 and field ψ_0 by

$$P_0 = \frac{n_{\mathrm{co}}}{2}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} |a_0|^2 \int_{A_\infty} |\psi_0|^2 \mathrm{d}A \tag{A1.5.114}$$

where the integration is over the infinite cross-section. Suppose we assume that this expression is invariant at every position along the taper, i.e. the total power remains fixed for all values of z. The field ψ_0 will then vary with the change in cross-section, i.e. $\psi_0 = \psi_0(z)$, and equation (A1.5.114) determines that the modal amplitude a_0 must also vary as

$$a_0(z) = \left(\frac{P_0}{\dfrac{n_{\mathrm{co}}}{2}\left(\dfrac{\varepsilon_0}{\mu_0}\right)^{1/2} \displaystyle\int_{A_\infty} |\psi_0(z)|^2 \mathrm{d}A \blacksquare} \right)^{1/2} = \left(\frac{P}{N_0}\right)^{1/2}. \tag{A1.5.115}$$

The denominator in the first expression is identical to the definition of the normalization N_0 in equation (A1.5.11), so on recalling the definition of the field of an *orthonormal mode* in equation (A1.5.12) it follows that

$$\hat{\psi}_m = \frac{\psi_m}{N_m^{1/2}}; \quad a_m \psi_m = P \frac{\psi_m}{(N_m)^{1/2}} = P \hat{\psi}_m \tag{A1.5.116}$$

In other words, the orthonormal fundamental mode describes the modal field everywhere along the taper if it is adiabatic but is a good approximation if the taper is approximately adiabatic.

A1.5.9.5 Mode coupling

In situations where there is a loss of power from a mode of a waveguide or fibre due to either an external influence, as is the case with multiple cores in Section A1.5.9.2, or to the longitudinal change in the profile or cross-sectional geometry due to tapering in Section A1.5.9.3, the variation in the power of the mode can be quantified using a set of *coupled-mode equations* [10]. An individual mode of any order propagates without change in its total power, field and propagation constant along a translationally invariant, loss-less waveguide or fibre.

If the waveguide or fibre is now subject to a length-dependent perturbation, then the mode generally will couple its power to all other bound modes and to the radiation field, in a multimode guidance situation, but only to the radiation field in a single-mode guidance situation, assuming an infinite cladding. For a finite-cladding cross-section, the radiation field comprises guided cladding modes together with a radiation field propagating into the medium beyond the cladding (normally air). Accordingly, only coupling between bound (core and cladding) modes will be included from hereon.

Forward and backward-propagating modes

The coupling of a particular mode, say the *j*th forward-propagating mode, is generally to all other forward-propagating modes and to all backward-propagating modes, including the *j*th backward-propagating mode. The *j*th forward-propagating mode in the weak-guidance approximation has a scalar transverse electric field E_j with the complete temporal-spatial dependence

$$E_j(x, y, z, t) = a_j(z)\psi_j(x, y)\exp[i(-\beta_j z - \omega t)] \tag{A1.5.117}$$

where $a_j(z)$ is the z-dependent modal amplitude, which accounts for the coupling of power to or from the *j*th mode. The *k*th backward-propagating mode has the same field as the *k*th forward-propagating mode, and its propagation constant has the same absolute value, but opposite in sign to that of the forward-propagating mode to account for the negative phase velocity, i.e.

$$E_k(x, y, z, t) = a_k(z)\psi_k(x, y)\exp[i(-\beta_k z - \omega t)] \tag{A1.5.118}$$

If $k = j$, this gives the form of the *j*th backward-propagating mode.

Mode normalization

If $\psi_j(x, y)$ is a solution of the scalar wave equation satisfying the usual boundary conditions, then any multiple $p\psi_j$ is also a solution, where p is an arbitrary constant. Thus the ψ_j in the above expressions are not uniquely specified, which in turn affects the value of the amplitude coefficients $a_j(z)$. To remove this potential ambiguity, it is convenient to replace the ψ_j and the ψ_{-k} by their orthonormal forms, i.e. by

$$\psi_j(x, y) \rightarrow \frac{\psi_j(x, y)}{\sqrt{N_j}}; \quad \psi_{-k}(x, y) \rightarrow \frac{\psi_{-k}(x, y)}{\sqrt{N_k}} \tag{A1.5.119}$$

where N_i and N_k are the respective mode normalization defined by equation (A1.5.11). Note that normalization is independent of the direction of propagation.

Coupled-mode equations

The coupled-mode equations comprise a set of linear, first-order differential equations that determine the z-dependence of the modal amplitude of each forward or backward-propagating mode. It is convenient to combine the z-dependent amplitude and phase of each mode by defining new coefficients $b_j(z)$ and $b_{-k}(z)$ such that (see chapter 27 of Ref. [4])

$$b_j(z) = a_j(z)\exp(i\beta_j z); \quad b_{-k} = a_{-k}(z)\exp(i\beta_k z) \tag{A1.5.120}$$

The coupled equations can then be expressed as

$$\frac{\mathrm{d}b_j}{\mathrm{d}z} - i\beta_j b_j = i\sum_k C_{jk} b_k; \quad \frac{\mathrm{d}b_{-j}}{\mathrm{d}z} + i\beta_j b_{-j} = -i\sum_k C_{-jk} b_k \tag{A1.5.121}$$

for the jth forward and backward-propagating modes, respectively, where the summation in k is over all forward and backward-propagating modes. The $C_{jk} = C_{jk}(x,y)$ *coupling coefficients*, defined by

$$C_{jk} = \frac{k}{2n_{\mathrm{co}}} \int_{A_\infty} [\bar{n}^2(x,y,z) - n^2(x,y)]\psi_j \psi_k \, \mathrm{d}A \qquad \text{(A1.5.122)}$$

where A_∞ is the infinite cross-section, $n^2(x,y)$ is the refractive index profile of the unperturbed waveguide or fibre, $n^2(x,y,z)$ is the refractive index profile of the perturbed waveguide or fibre, and the ψ values are orthonormal.

Applications of coupled-mode equations

The set of coupled-mode equations is useful for solving a range of practical propagation problems where there is a significant exchange of power between modes. Such an analysis requires a set of propagation conditions along the fibre, which does not change or only changes very slightly so that the field and propagation constant of each mode remains constant. Examples where coupled-mode equations offer a method of solution include multiple-core fibres and waveguides, as discussed in Section A1.5.9.2, Bragg reflection gratings, and long-period gratings.

Coupled local mode equations

Problems in which the translational invariance of the modes no longer holds, such as well-tapered fibres, waveguides or devices, are not appropriate for a coupled-mode equation analysis. In these cases, an alternative analysis based on an analogous set of coupled local-mode equations can be used (see chapter 28 of Ref. [4]).

In practice, however, problems involving tapered fibres, waveguides and devices, based on single or two-mode propagation, generally have a sufficiently slow rate of taper that coupling between local modes may be negligible so that each local mode propagates approximately adiabatically and almost conserves its power.

References

[1] Hecht J 1999 *City of Light—The Story of Fibre Optics* (Oxford: Oxford University Press)

[2] Kao K C and Hockham G A 1966 Dielectric-fibre surface waveguides for optical frequencies *Proc. IEE* **113** 1151–1158

[3] Snyder A W and Love J D 1975 Reflection at a curved dielectric interface—electromagnetic tunnelling *IEEE Trans Microwave Theory Tech* **23** 134–141

[4] Snyder A W and Love J D 2000 *Optical Waveguide Theory* (Dordrecht: Kluwer Academic Publishers)

[5] Besley J A and Love J D 1997 Supermode analysis of fibre transmission *Proc IEE* **144** 411–419

[6] Ladouceur F and Love J D 1996 *Silica-based Buried Channel Waveguides & Devices* (London: Chapman & Hall)

[7] Snyder A W 1969 Asymptotic expressions for eigenfunctions and eigenvalues of dielectric or optical waveguide *IEEE Trans Microwave Theory Tech* **17** 1130–1138

[8] Gloge D 1971 Weakly guiding fibers *Appl Optics* **10** 2252–2258

[9] Born M and Wolf E 1980 *Principles of Optics* Section 3.2.1

[10] Marcuse D 1974 *Theory of Dielectric Optical Waveguides* (San Diego: Academic Press)

A1.6

Introduction to lasers and optical amplifiers

William S Wong, Chien-Jen Chen and Yan Sun

A1.6.1 Introduction

Although Albert Einstein did not invent the laser, his work laid the foundation for its development. In 1917, Einstein was the first to explain how radiation could induce, or stimulate, more radiation when it interacts with an atom or a molecule [14]. A few years later, Richard Tolman discussed stimulated emission and absorption in his paper [65], realizing the important fact that stimulated emission is coherent with the incoming radiation. In other words, the electric dipoles in the atoms oscillate with the incoming photons, which in turn, re-radiate photons that have a fixed phase relationship with incoming photons. If the re-radiated photons are in-phase with the incoming ones, they add constructively to amplify the incoming photons. Thus, the general idea of coherent amplification via stimulated emission was understood since the 1920s. However, it was not until the 1950s when the concept of the 'maser', which is an acronym for microwave amplification by simulated emission of radiation, was developed and demonstrated by Charles Townes and his coworkers at Columbia University [22, 23]. They directed excited ammonia molecules into a cavity whose resonance frequency is tuned to the 24 GHz transition frequency of ammonia [66]. A sufficient number of these excited molecules will initiate an oscillating microwave field in the cavity, part of which will be coupled out of the cavity (see figure A1.6.1). It is interesting to note that maser operation was first demonstrated in the microwave region. Since the spontaneous radiative lifetime is inversely proportional to the third power of the transition frequency, at microwave transition frequencies, the radiative lifetime of the ammonia molecules is about 1×10^{12} longer than it would be at optical frequencies, which allows the system to achieve population inversion easily with a reasonable amount of pump power.

Following the invention of the maser, researchers attempted to create a laser by extending the maser action to optical frequencies, where 'laser' is an acronym for light amplification by simulated emission of radiation. Schawlow and Townes published a paper to explore the possibility of laser action in the infrared and visible spectrum [50]. Although many researchers at the time speculated that a system containing alkali vapours would be the most likely candidate that could be made to oscillate at optical frequencies, they were surprised to learn that a system involving optical pumping of chromium ions in ruby, invented by Maiman [37], was the first to produce coherent optical radiation. His laser consisted of a ruby crystal surrounded by a helicoidal flash tube enclosed within a polished aluminum cylindrical cavity (see figure A1.6.2). When pumped by a very intense pulse of light from a flash lamp, the ruby laser operated in pulsed mode at 694 nm.

Great progress has been made since the 1960s. Today, lasers have become ubiquitous in our lives. They are used in a wide range of applications in communications, medicine, consumer electronics, and military systems. Although it has been fewer than 50 years since the laser was first invented, we have made, and will continue to make, great strides in advancing the state of the art in lasers.

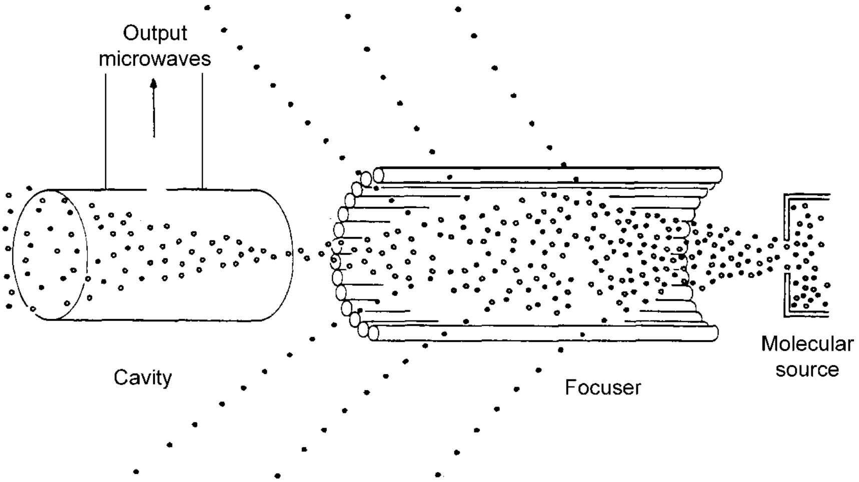


Figure A1.6.1. The ammonia maser. Molecules diffuse from the source into a focuser where the excited molecules (shown as open circles) are focused into a cavity and molecules in the ground state (shown as solid circles) are rejected. A sufficient number of excited molecules will initiate an oscillating electromagnetic field in the cavity, which is emitted as the output microwaves (from [66]).

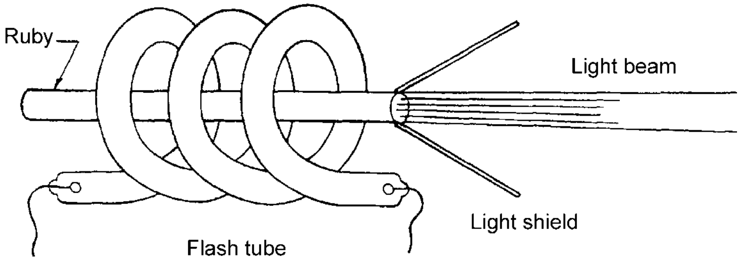


Figure A1.6.2. Schematic diagram of a ruby laser. When the gas flash tube is activated, electromagnetic oscillations occur within the ruby rod. Some of the visible radiation is emitted in a beam through one partially reflecting end of the rod (from [66]).

A1.6.1.1 A two-level atomic system

Before we explain the operation of a laser, we need to understand how radiation interacts with matter. Atoms, molecules, and solids have specific energy levels that are determined by the laws of quantum mechanics. A photon may interact with an atom if its energy matches the difference between two of the energy levels.

Let us focus our attention on two energy levels, E_1 and E_2 of an atom. If we assume that $E_2 > E_1$, we can refer to the levels 1 and 2 as lower and upper levels, respectively. When the energy of the photon $h\nu$ matches the atomic energy-level difference, i.e. $h\nu = E_2 - E_1$, the photon interacts with the atom where processes such as spontaneous emission, absorption, and stimulated emission take place.

Spontaneous emission of a photon takes place when the atom transitions from level 2 to level 1 without any external excitation (figure A1.6.3). The energy difference is released in the form of a photon with energy $h\nu = E_2 - E_1$. The rate of this spontaneous transition, which is independent of the number

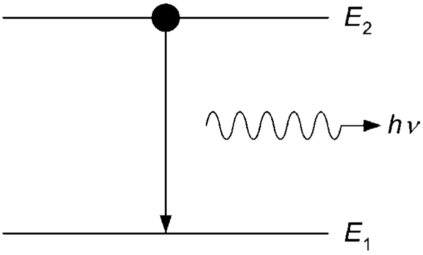


Figure A1.6.3. The process of spontaneous emission.

of photons that is present, is the reciprocal of the spontaneous radiative lifetime of the 2 → 1 transition. Typical values of radiative lifetimes of specific transitions range from picoseconds to several minutes.

In the presence of pumping, the atom will transition from the lower level to the upper level. For example, in the He–Ne laser, the He atoms, initially excited by collisions with electrons, transfer the excitation to the Ne atoms. In addition, in the presence of a photon with energy $h\nu = E_2 - E_1$, the atom in the lower level will also transition to the upper level, while absorbing the photon at the same time (figure A1.6.4). It is an induced transition since the rate of absorption is *proportional* to the intensity of the radiation. As an example, cesium atomic clocks, which utilize the principle of absorption, have the ability to measure time with an accuracy as good as 1 s in over 20 million years. Cesium is the best choice of atom for such a measurement, because all of its 55 electrons, except the outermost one, are confined to orbits in stable shells of electromagnetic force. The energy difference $E_2 - E_1$ is attributed to the cesium atom's outermost electron transitioning from a lower to a higher orbit. After tuning a microwave beam to the resonant absorption frequency (9,192,631,770 Hz) of a collection of cesium atoms, one measures the resulting cycles of oscillations in the microwave signal to establish a time/frequency standard.

If the atom is in the upper level in the presence of radiation, it will make a downward transition to the lower level while emitting a photon of energy $h\nu$ (figure A1.6.5). The emission of the new photon is

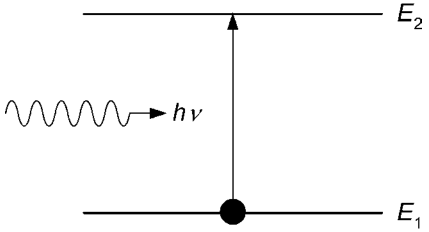


Figure A1.6.4. The process of stimulated absorption.

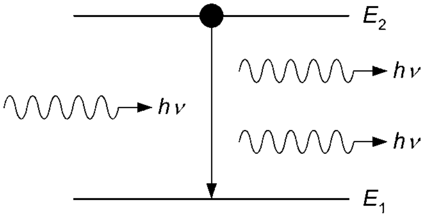


Figure A1.6.5. The process of stimulated emission.

induced, or stimulated, by the incident photon. The new photon also has the same frequency, polarization, and phase as the incident photon. In an optical amplifier, this process of stimulated emission takes place continuously along the length of the gain medium, resulting in a large number of output photons that are replicas of the input photons. In general, in a spontaneous process, such as spontaneous emission, each atom acts independently to produce a noise-like output. On the other hand, in a stimulated process, the atoms in the medium act collectively and oscillate in phase, which makes coherent amplification possible.

Although no photons are involved, nonradiative decays are also an important process for understanding the operation of a laser. If an atom is initially in the upper energy level, it can decay to the lower energy level without emitting a photon. Instead, the energy-level difference $E_2 - E_1$ appears in a phonon, where either the rotational energy or the translational energy of the material system increases, resulting in heating effects.

At thermal equilibrium, the ratio of the atoms in levels 1 and 2 is given by the Boltzmann factor:

$$N_2/N_1 = \exp[-(E_2 - E_1)/k_B T], \tag{A1.6.1}$$

where N_1 and N_2 are the number of atoms in the lower and upper levels, respectively, $k_B \approx 1.38054 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ is the Boltzmann constant, and T is the temperature in kelvin. Since $E_2 > E_1$ by assumption, the above equation tells us that at thermal equilibrium, $N_2 > N_1$; in other words, there are more absorbers than emitters. Therefore, at a normal temperature where $T > 0$, the two-level system always absorbs more incoming radiation with energy $h\nu$ than it will emit. In order to have a net amplification of radiation, it is necessary to achieve a population inversion, where $N_2 > N_1$. Interestingly, the corresponding system is said to have a 'negative temperature' since T will have to be a negative quantity in kelvin in order to satisfy equation (A1.6.1).

A1.6.1.2 Multi-level laser systems

Although, 'laser' is an acronym for 'light amplification by stimulated emission of radiation', lasers are actually optical oscillators (generators or sources of light) and not optical amplifiers (devices for increasing the strength of an input lightwave signal). Despite the variety in the types of lasers, nearly all lasers possess the following three essential elements:

(a) a lasing medium, which can be a solid, liquid, gas, semiconductor, or others;

(b) a pumping process that excites the atoms in the medium into a higher energy level; and

(c) an optical resonator to enable the light to bounce back and forth.

If the amount of coherent amplification inside the resonator exceeds the internal loss, coherent laser oscillation will take place.

It is impossible to achieve continuous laser oscillation in a two-level atomic system. Imagine that we are trying to excite as many atoms into the upper level as possible. While doing so, we are also increasing the rate of stimulated absorption. At best, both upper and lower levels in the two-level system will be populated equally, failing the requirement that we need to have a population inversion ($N_2 > N_1$). To overcome this problem, most laser systems utilize three or more energy levels. A three-level system is shown in figure A1.6.6. Atoms are excited from level 1, the ground state, to level 3, which decay quickly to level 2, also known as a metastable state. Provided that there is a population inversion, stimulated emission, which takes place from level 2 to level 1, will dominate. This condition is met if the nonradiative lifetime from level 3 to level 2 is short, so that it will not give the pumping process a chance to re-populate the ground state. The ruby laser invented by Maiman is an example of a three-level laser system.

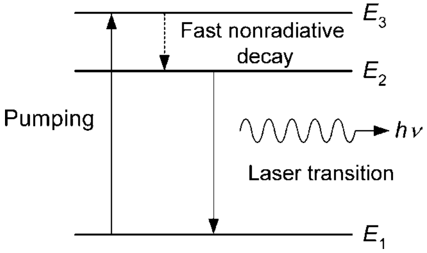


Figure A1.6.6. A typical three-level laser.

A1.6.1.3 Laser cavity

A major element of a laser, in addition to the gain medium, is the laser cavity, which provides optical feedback so that laser oscillation can be achieved. It works like an electronic oscillator where stored energy, amplification, feedback, and output coupling are all engineered for oscillations to take place. It is straightforward to form a laser cavity. In fact, in the simplest form, two parallel mirrors facing each other form the most basic laser cavity—the Fabry–Perot cavity [27, 49, 51, 74]. Lightwave bounces back and forth upon both mirrors, gaining energy each time it passes through the gain medium. Typically, one mirror is highly reflective and the other is partially transmitting to provide output coupling. The gain, loss, and coupling ratios of the mirrors determine the output power of a laser.

Since light is circulating inside the cavity, the phase of the reflected light, after a complete round trip, needs to match that of the original light so that a constructive interference or a standing wave can form. This condition is satisfied only at certain frequencies and these resonant frequencies are referred to as the longitudinal (or axial) modes of the laser cavity. For a cavity with a fixed length, the longitudinal modes are separated by a constant frequency that is equal to the inverse of the cavity round trip time. Within the gain bandwidth of the gain medium of a laser, there may exist multiple longitudinal modes. The lasing threshold is reached when the gain medium is pumped, in which the total round-trip gain of the mode (closest to the gain peak) equals the total round-trip loss. When the threshold is exceeded, the laser oscillation starts in this axial mode. Depending on the property of the gain medium, when the pump power increases, a homogeneously broadened gain medium, which maintains the same gain shape when saturated, can sustain one stable lasing axial mode only, while an inhomogeneous broadened medium can have multiple modes co-existing in a laser.

Similar to a RLC electronic resonator, the quality factor Q of a laser cavity is defined as 2π times the ratio of the stored energy to the energy loss per cycle. Since an optical cycle is very short (a few femtoseconds), energy loss per cycle is relatively small and the quality factor of an optical cavity is normally much greater than that of an electronic resonator.

An important method to generate short laser pulses is to 'switch' the cavity quality factor artificially from very low to normal so that the stored energy in the gain medium can be released in a short time period, otherwise known as the Q-switching method. Lowering the cavity Q can be achieved by increasing the cavity loss or decreasing the feedback from the mirrors. The stored energy, in the form of large inversion in the gain medium, grows as pump power is applied, but the laser oscillation does not happen because of the high cavity loss or the lack of optical feedback. When the cavity Q is restored, the laser threshold is exceeded and the lightwave inside the cavity experiences a large optical gain due to the large inversion and builds up rapidly. The huge optical energy then quickly consumes the large population inversion causing the optical power starts to decrease. In addition, when the cavity Q is lowered at the same time, the optical power diminishes further as the laser is operating below its threshold. The cycle is then repeated for subsequent pulse generation. The Q switching technique is widely applied on all types of lasers, from solid-state lasers, gas lasers, to semiconductor lasers, to produce short pulses.

In addition to the Fabry–Perot cavity, there are different ways to construct a laser cavity. One can replace the planar mirrors with concave or convex spherical mirrors to achieve better confinement of the optical energy inside the cavity [4]. An important consideration of a laser cavity is whether a stable mode exists that corresponds to the mirror geometry. This problem can be approached in a setting of the paraxial wave equation, which describes wave propagation in the paraxial limit, and its eigen solutions, i.e. the Hermit–Gaussian modes. Another approach is ray tracing, which is described by the ABCD matrices. Both methods lead to the same stability criterion written explicitly as $0 \leq g_1 g_2 \leq 1$ [27, 49, 51, 74], where the g parameter of a mirror is defined as $g = 1 - L/R$, where L is the distance between the two mirrors and R is the radius of curvature of the mirror.

As described by the same analysis, higher order Hermit–Gaussian modes, or the transverse modes, can exist in a laser cavity to compete for gain, degrade mode profile, and cause output power fluctuation. Since the higher order transverse modes have larger spatial profiles, they can be eliminated by restricting the active region of the gain media to fit only the fundamental mode. In some high gain lasers, unstable cavities, in which off-axis light diverges, can be used as they have advantages of a larger spatial profile, more efficient gain, ease of alignment, and single modalness.

There are other types of laser cavities, such as a ring cavity consisting of a length of optical fibre. In semiconductor lasers, cleaved facets can be used as mirrors; in addition, linear cavities formed using distributed Bragg reflectors are also common. Depending on the available technology and requirements, laser cavities are chosen to suit their purposes.

A1.6.2 Specific types of lasers

A1.6.2.1 Solid-state lasers

Since the invention of the very first ruby laser in 1960s [37, 38], solid-state lasers, which are composed of gain media in their solid phase by definition, remained as one of the most important and versatile lasers for academia and industry. When pumped to threshold population inversion levels, the three-level or four-level solid-state gain media provide sufficient gain to compensate for losses in the laser cavity. The laser cavity forms a resonator, analogous to the electronic version of an oscillator, which determines the output power, operation modes, and the beam profile of the solid-state laser.

Generally, a solid-state gain medium consists of a crystalline (e.g. YAG) or an amorphous (e.g. glass) host doped with a certain ion (e.g. several percent of Cr^{3+}, Nd^{3+}, Er^{3+}, etc) that acts as the active lasing material. Most commercially available solid-state lasers are pumped by gas discharge flash lamps (e.g. krypton or xenon) or by semiconductor lasers for greater reliability and efficiency. The choice of the gain media depends on lasing wavelength, optical properties (birefringence, dispersion, optical nonlinearity), and mechanical properties (material strength, thermal conductivity). Important solid-state gain media include Nd^{3+}-doped YAG (Nd^{3+}:YAG), Nd^{3+}:glass, Ti^{3+}:sapphire, and colour-centre lasers, etc.

The Nd^{3+}:YAG laser is a four-level laser and is typically operating at the 1.064 μm wavelength range with a narrow linewidth of a few nanometers [18, 35]. Its good thermal conductivity and temperature-insensitive threshold make it suitable for both pulsed and continuous-wave operations. Typically, it is pumped optically with flash lamps or laser diodes, and water-cooled in high output power applications. It is popular for various laboratory applications such as pumping another laser or being frequency-doubled to provide green light. The energy-level diagram of the Nd^{3+}:YAG laser and its emission spectrum are shown in figure A1.6.7.

Compared with Nd^{3+}:YAG, the Nd^{3+}: glass is easier to manufacture and process. It can be made into a larger size gain medium to achieve higher output powers. Depending on the host glass types, it

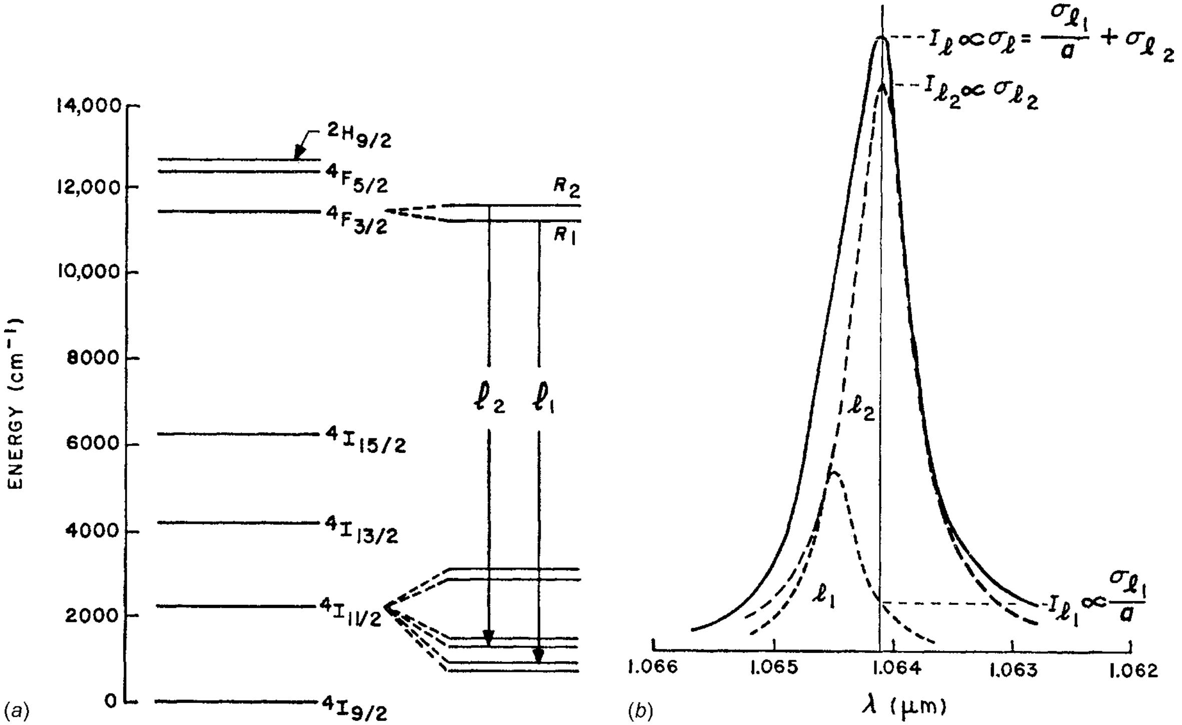


Figure A1.6.7. (*a*) Energy-level diagram of Nd^{3+}:YAG (from [35]), (*b*) Spontaneous-emission spectrum of Nd^{3+}:YAG near 1.064 μm at room temperature. The two Lorentzian lines contributing to the laser transition are shown by dashed lines (from [35]).

typically operates at around 1.06 μm with a wider linewidth of tens of nanometres [34, 54]. Using a large size Nd^{3+}:glass gain medium and the *Q*-switch technique, one can generate pulses with terawatt peak power. However, due to its low thermal conductivity, the Nd^{3+}:glass laser operates at a low repetition rate (10 pulses per second). For comparison, we show the fluorescent emission spectra of Nd^{3+} in various glass hosts in figure A1.6.8.

Among the solid-state gain media, Ti^{3+}:sapphire is characterized by its broad gain bandwidth from 0.66 μm to greater than 1.0 μm (more than 100 THz of bandwidth when modelocked) [43]. It is used in tunable continuous wave (CW) lasers for broad tuning ranges or in mode-locked configuration to produce ultra-short pulses (down to several femtoseconds). The Ti^{3+}:sapphire lasers are often pumped by other lasers such as laser diodes or argon ion gas lasers due to its short excited state lifetime (3.2 μs).

Two common operation modes of solid-state lasers are the CW mode and the pulsed mode. In the CW operation, solid-state lasers generate monochromatic, highly coherent, and high-intensity light. A widely used mechanism to produce pulses in solid-state and other lasers is the *Q*-switching method [68, 69]. By changing the cavity *Q* mechanically, electrically, or optically, the stored energy in the laser gain medium can be released with a short time period to generate a pulse (in the order of nanoseconds) with a high peak intensity [51, 67, 74] (see also A1.6.1.3 for a description of *Q* switching). Short pulses provide advantages in manufacturing and in medical applications such as micro-machining/micro-fabrication and laser ablation since the size of the heat-affected zone can be reduced. As the demand on shorter pulse duration and higher peak intensity increases, different techniques such as mode-locking techniques [26, 53] and chirped pulse amplification (CPA) techniques [60] were applied on solid-state

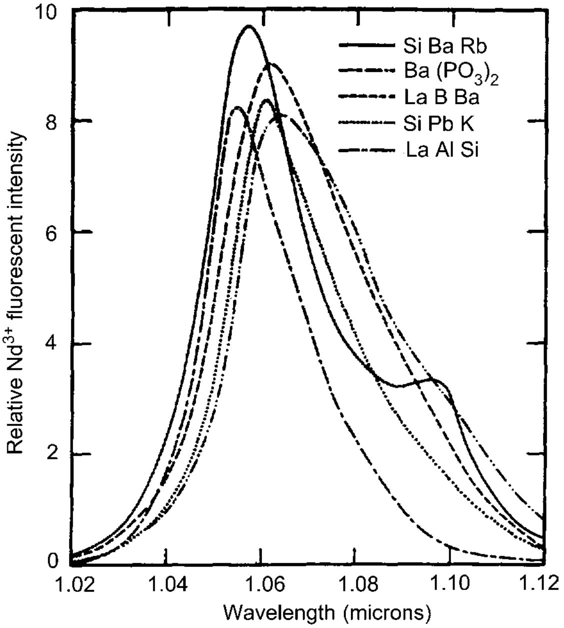


Figure A1.6.8. Spontaneous-emission spectrum of Nd^{3+} in various glass hosts near 1.06 μm (from [54]).

lasers to generate ultra-short pulses (down to sub-picosecond ranges). In fact, solid-state lasers, due to their wide bandwidth and excellent optical properties, generated many of the record-breaking short pulses. These ultra-short pulses are suitable for applications in the study of ultra-fast phenomena, spectroscopy, and telecommunication.

A1.6.2.2 Gas lasers

As another important family of lasers, gas lasers were fast growing in the laser industry due to their simple pumping schemes and the wide availability of gain media. In gas lasers, the population inversion is typically created by electric discharge, which is relatively simple to construct and operate. Although the gain media are in the gaseous phase, they can be made up of neutral atoms, ions, or gas molecules. As examples, we will briefly introduce the three most popular gas lasers: the He–Ne laser, the Argon ion laser (Ar$^+$), and the CO$_2$ laser.

The He–Ne laser, one of the best known lasers, operates at the famous 632.8 nm wavelength [5, 31]. The gain medium consists of a mixture of two noble gases, He and Ne, with a population ratio of about 10 to 1. The He atoms are first excited to higher energy levels by the electrical discharge and then pass energy to the Ne atoms through inelastic collisions. This is possible due to similar energy levels between the He and the Ne atoms (see figure A1.6.9). The excited Ne atoms provide gain for the laser operation in the He–Ne lasers, which also radiate at wavelengths of 1.15 μm and of 3.39 μm, in addition to the 632.8 nm. In fact, optical filtering in the laser cavity is often applied in the He–Ne lasers to reduce the internal gain in the infrared since the optical gain at 632.8 nm is relatively low. The He–Ne lasers were widely used in optical alignment, survey, and bar-code scanning, etc, before the advent of inexpensive laser diodes.

The Ar$^+$ lasers are capable of producing high power in the visible wavelength range. Hundreds of wavelengths could exist in an Ar$^+$ laser cavity; however, the 488.0 and 514.5 nm are two of the most

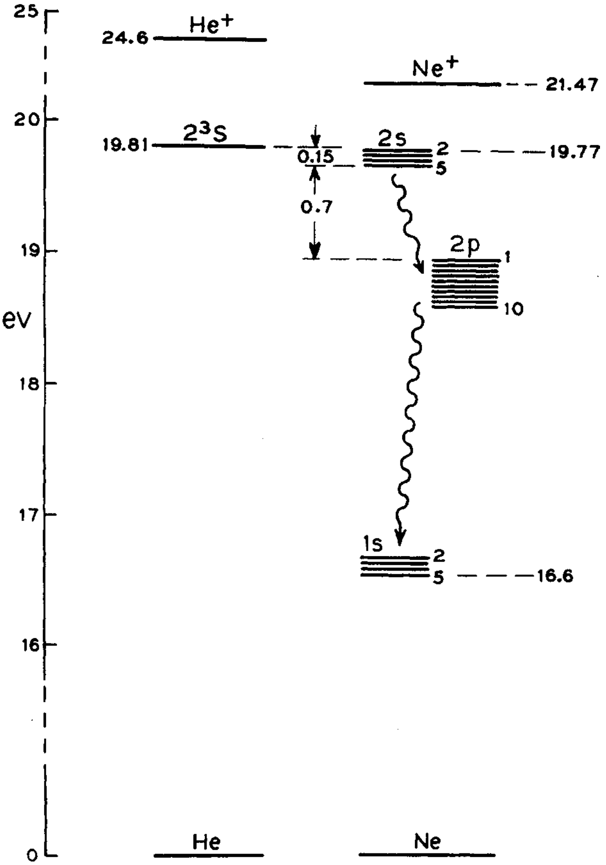


Figure A1.6.9. Energy level diagram of the He and Ne atoms (from [31]).

prominent wavelengths [6, 21]. The pumping mechanism in Ar^+ lasers is complicated, including multiple collisions between electrons, between Ar atoms, and between Ar^+ ions. Since the active laser medium consists of Ar^+ ions, whose ground state level is 16 eV higher than that of the Ar atoms, much of the pump power is wasted in providing ionization of the Ar atoms. Because of the ground level difference, which does not contribute to the laser oscillation, the Ar^+ laser threshold current density is typically high and the efficiency of the laser is low. Due to the low efficiency, bulky water-cooling subsystems are used to cool high-power Ar^+ lasers.

The CO_2 lasers are gas lasers and are well known for their high efficiency and high output powers [45, 46]. In contrast with the He–Ne laser and the Ar^+ laser, in which transitions in electron energy levels provide laser gain in the cavity, CO_2 lasers operate on molecular vibration modes. The vibration modes are represented by three quantum numbers (n_1, n_2, n_3) where the ground state is at (0, 0, 0). Typically, N_2 molecules and He atoms are also present in the CO_2 laser cavity. The N_2 molecules store the energy from the electrical discharge in the fundamental vibration mode and then transfer the energy to the asymmetric vibration (0 0 1) mode through collision since both vibration modes have similar energy levels. The CO_2 laser gain is then produced from the transition of the asymmetric vibration mode to lower level vibration modes, including wavelengths of 10.6 and 9.6 μm. The He atoms, moving at higher speed, effectively reduce the lifetime of the vibration modes at lower levels of the laser and remove heat. This process makes population inversion possible and increases the overall efficiency. Due to its high output power, the CO_2 laser is used in industries such as metal processing and machining.

A1.6.2.3 Semiconductor lasers

A semiconductor (e.g. silicon, germanium, gallium arsenide) is a material whose electrical properties are between those of a conductor and an insulator. The electrons in a semiconductor are found in bands that are separated by a band gap—the lower band is the valence band while the upper band is called the conduction band. For example, GaAs has a bandgap of 1.424 eV. Furthermore, by doping intrinsic semiconductor materials with impurities, one can make n-type or p-type materials that have more or fewer negative current carriers.

Shortly after the invention of the first laser, the first semiconductor laser was demonstrated independently in 1962 by four research groups in the United States [25, 30, 44, 47]. These early devices are homojunctions operated at liquid nitrogen temperatures. When a current (defined as the flow of *positive* charges) is injected so that it flows through the junction from a p-type material into an n-type material, electrons from the n-type material will recombine with the holes from the p-type material, releasing a form of energy known as recombination energy. In an indirect-bandgap material such as silicon, this energy is released as vibrational energy and heat; on the other hand, in a direct-bandgap material such as gallium arsenide, radiation is emitted whose frequency is a function of the bandgap energy. In the absence of optical feedback, this device functions as a light-emitting diode (LED) where its output consists of incoherent spontaneous emission. However, feedback from reflective surfaces is made possible using the cleaved facets. Since the refractive index of a semiconductor is usually greater than 3.0, the reflectivity of each cleaved facet can be as high as 25% without any coatings.

By the late 1970s, semiconductor lasers are able to operate in CW mode at room temperature. These lasers utilize a double heterostructure (figure A1.6.10) to improve the confinement of light in the active region, resulting in lower threshold currents [2, 28]. Later, the incorporation of thin quantum wells into the heterostructure offers additional advantages such as low threshold current density, high efficiency, and high differential gain. Figure A1.6.11 shows that the threshold current density has been dramatically reduced by a factor of 10,000 since the early 1960s.

The technology of semiconductor lasers has been advanced to a point where, without realizing it, an average person owns at least a few of them. Examples are lasers used in CD or DVD players, in laser printers, and in laser pointers. In addition, in fibre optics communications, semiconductor lasers are used as transmitters and as EDFA and Raman pump lasers.

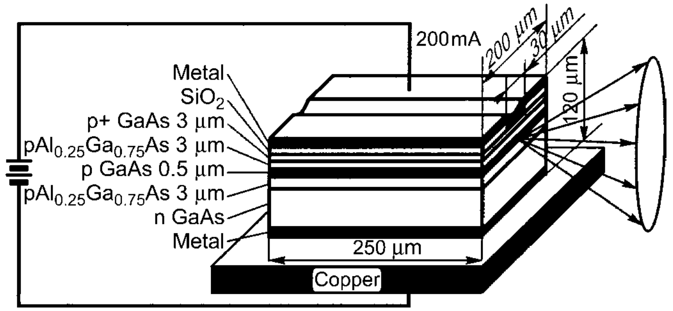


Figure A1.6.10. Schematic view of the structure of the first injection double-heterostructure laser operating in the CW regime at room temperature (from [2]).

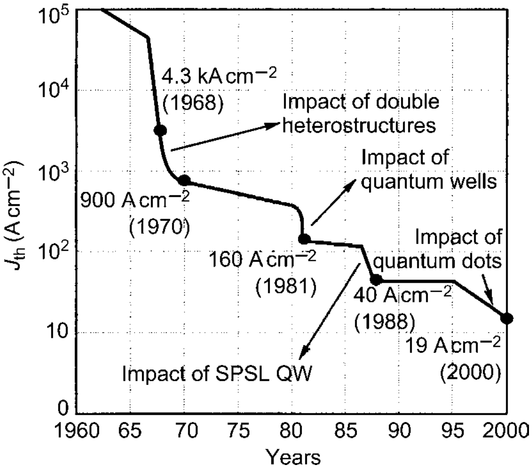


Figure A1.6.11. Improvements of the threshold current of semiconductor lasers over the last 40 years (from [3]).

A1.6.2.4 Short-pulse lasers

When the gain medium of a Fabry–Perot laser is homogenously broadened, only one single longitudinal mode oscillates. On the other hand, if the gain medium is inhomogenously broadened, for example, via spatial hole burning or spectral hole burning, a few longitudinal modes will oscillate, provided that their round-trip gain is greater than unity. These longitudinal modes are spaced evenly in the frequency domain with frequency spacing $\Delta v = c/2nL$, where n is the average refractive index inside the Fabry–Perot cavity, and c is the speed of light in vacuum. If we are able to lock these modes together such that their phases are fixed relative to one another, we can create a train of laser pulses at the output of a laser. This technique, known as modelocking, can be initiated actively using a time-varying amplitude modulation, or it can be triggered using a saturable absorber in the cavity whose transmissivity decreases with increasing optical intensity. Very short optical pulses can be generated from modelocked lasers because the electric field from the various longitudinal modes are coherent—they interfere constructively at the peak of the pulse while cancelling one another at the wings of the pulse.

To capture fast-moving images on film, a photographer uses a fast shutter setting. Since the speed of mechanical shutters is limited to milliseconds, improvement is made to capture faster events by illuminating the object with a stroboscope. As one of the pioneers in strobe-light photography, Edgerton captures the dramatic image of a rifle bullet piercing an apple in one of his high-speed photographs (figure A1.6.12). The photos are taken in a room with the lights turned off. When the bullet is fired, its shock wave is detected by a crystal microphone, which then triggers the strobe light. Edgerton's technique enables him to freeze a microsecond event onto a photographic negative. With the advancement of ultrafast laser research, nanosecond, picosecond, and subsequently, femtosecond laser pulses are available for scientists to 'freeze' and study physical and chemical processes in a very short time scale. In fact, some used to believe that the various chemical and biological processes, such as the breaking of chemical bonds and vision, were slow processes and did not occur in the femtosecond time scale, until photochemistry experiments using short laser pulses proved otherwise. These results can be explained by the fact that, since molecular motions occur over very short distances, they can be very fast.

Given the existence of the various techniques to produce femtosecond pulses, how does one measure the duration of these pulses? In general, to measure a fast event, one needs an even faster event in order to capture it, which is not often possible since the laser pulse in interest might be the

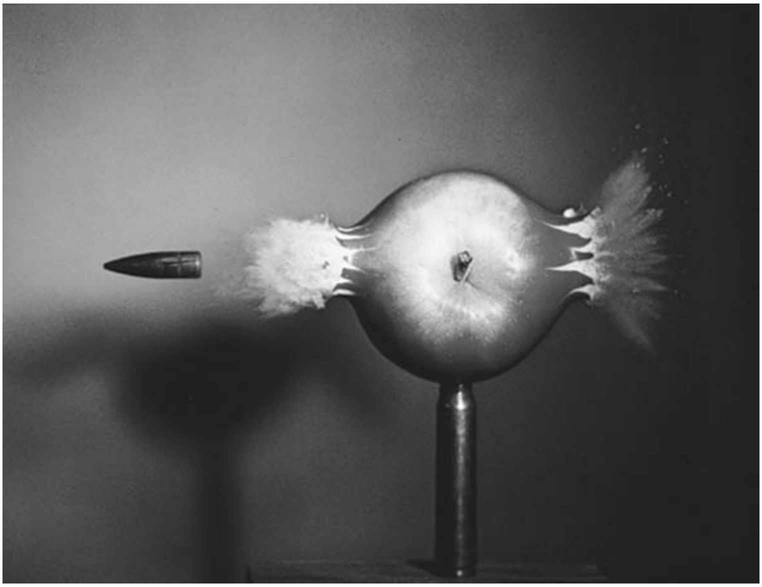


Figure A1.6.12. A microsecond exposure of a bullet travelling 2800 ft s^{-1} while piercing an apple (from http://www.geh.org/taschen/htmlsrc4/m199603470007_ful.html).

shortest event available. Likewise, researchers working in scanning-tunneling microscopy face a similar problem when they want to establish the spatial resolution of their system—the fine tip they use to probe the sample is often the finest man-made object that is available to them. In 1967, Weber found a partial solution; he suggested measuring the laser pulse width with the pulse itself by performing an intensity autocorrelation [71]. The optical pulse is split into two identical pulses using a beam splitter. The two pulses are then focused onto a nonlinear crystal that is capable of generating a second harmonic. The second harmonic generated will be collected in a photomultiplier tube while the temporal delay between the two pulses is varied. For example, the largest amount of signal is obtained when the two pulses overlap temporally. This operation is identical to performing an autocorrelation function mathematically, from which the approximate pulse duration can be deduced. Although it is a clever technique, the autocorrelation method does not yield complete details about the intensity profile of the pulse. Nor does it provide any information regarding the phase or the chirp of the optical pulse.

Ideally, one would like to get pulse information in both time and frequency domains, which is usually referred to as a spectrogram. It is analogous to the musical score for a symphony, informing musicians which notes to play at a given time. It was not until 1993 that researchers developed a novel method to retrieve to obtain the spectrogram of an optical pulse. Using a technique called frequency-resolved optical gating [32], or FROG, they measured the optical spectrum of the autocorrelation and then applied a phase retrieval algorithm to obtain the final spectrogram. Their algorithm works because knowledge of only the magnitude of a two-dimensional Fourier transform of a function of two variables uniquely determines the function (both phase and magnitude), provided that the function is well behaved. Figure A1.6.13 shows that both the intensity autocorrelation and the interferometric autocorrelation cannot distinguish positively chirped pulses from negatively chirped ones. On the other hand, the spectrogram extracted using FROG resolves this ambiguity as it contains complete amplitude and phase information about the optical pulse.

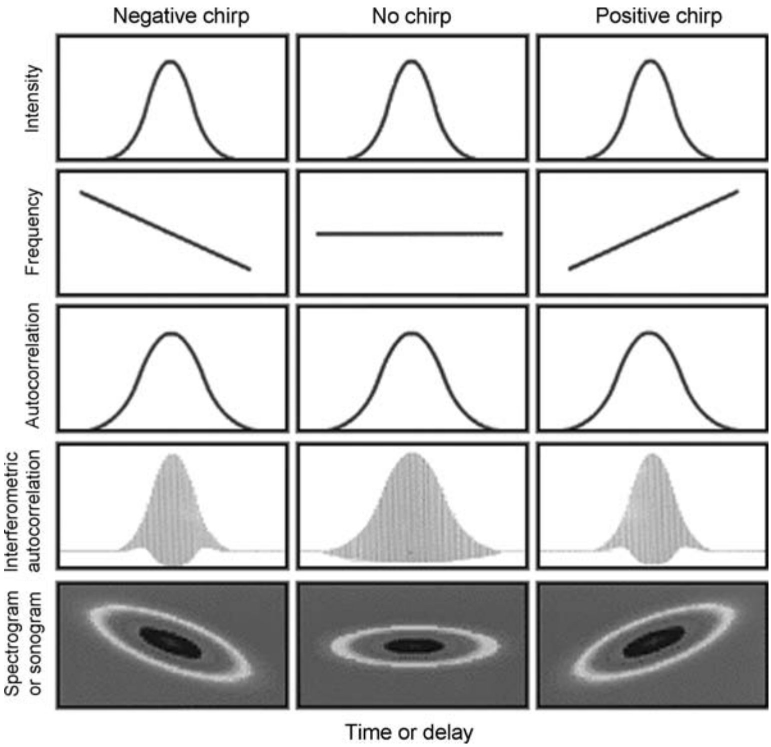


Figure A1.6.13. The intensity versus time, the frequency versus time, the intensity autocorrelation versus delay, the interferometric autocorrelation versus delay, and spectrograms (or sonograms) of negatively chirped, unchirped and positively chirped Gaussian-intensity pulses. In the spectrograms, the vertical axis is frequency and the intensity is colour-coded. Note that the autocorrelation and interferometric autocorrelation cannot distinguish positive from negative chirp, while the spectrogram and sonogram can (from [70]).

A1.6.3 Optical amplifiers

A1.6.3.1 Basics of optical amplifiers

Optical amplifiers are used to boost the optical power of the input optical wave. Optical amplifiers are essentially 'single-pass' lasers, or 'lasers' without the two (reflective) mirrors. Just as in the case of lasers, the gain medium is 'inverted' and provides stimulated emission when there is an input optical wave. In principle, all types of lasers can be converted into optical amplifiers. For simplicity, we will use optically pumped amplifiers in the following discussions. In such amplifiers, three optical waves are present, the pump wave that provides the inversion in the gain medium, the input optical wave that is the input signal to be amplified, and the output wave that has been amplified in the amplifier.

There are different applications for various types of optical amplifiers. Accordingly, the parameters of interest also vary. For example, the gain is an important parameter for laser fusion, while noise is essential for telecom applications. In general, the three commonly used parameters are gain, output power, and noise figure.

Gain is defined to be the ratio between the output power and input power, where it is usually measured in decibels. Unity gain, or 0 dB, means no gain or loss. On the other hand, most practical optical amplifiers operate in gain regions much higher than 1. A typical gain curve is shown in

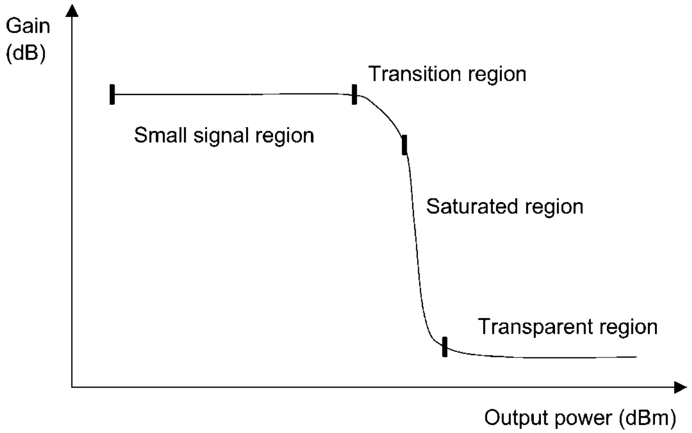


Figure A1.6.14. Gain versus output power for a fixed pump power.

figure A1.6.14 as a function of output power for a fixed pump power. The gain is also commonly plotted as a function of input power. There are four regions in this saturation curve: the small signal region, the transition region, the saturated region, and the transparent region.

In the small signal region, the input power is low, and the gain remains constant when input power changes. The inversion in the gain medium is determined only by the amount of pump power and is independent of the weak input optical power. An advantage for this operation region is that the amplifier's gain value remains constant. However, the photon conversion efficiency from pumps to signals is low. Most of the pump photons are not converted into signal photons, but into spontaneous emission or heat or other forms of energy.

As input power increases into the transition region, the interaction between the signal and the gain medium gets stronger. In this region, the signal starts to deplete the gain medium and the inversion level starts to drop. This causes the gain to decrease when input power increases.

When the input power is high, the amplifier works in the saturated region where the gain drops sharply with increased input power. The inversion level is low and the photon conversion efficiency from the pump wave to the signal wave is high. This is a common operation region for amplifier design in order to make efficient use of the pump power. In this highly saturated region, the output power changes very little as input power varies, as shown in figure A1.6.14.

In the limit of high input power, the amplifier moves into the transparent region where the pump is relatively weak and the input optical wave essentially bleaches through the medium. There are approximately equal number of atoms in the upper level and in the lower level. No practical amplifiers are designed to work in this region.

During the amplification process, spontaneous emission from the gain medium may add to the optical signal wave. Noise figure is the parameter to measure how much noise is added in the amplifier. A low noise figure is important for telecom applications, as we desire to minimize the degradation of the signal-to-noise ratio when a signal is amplified. The quantum limit for a practical optical amplifier is 3 dB [1]. Under the assumption of high-gain operation for the amplifier, high inversion level allows for low noise figure while low inversion level yields high noise figure.

There are other parameters that need to be considered when designing an amplifier, such as reliability, size, cost, etc. We will not discuss the details here as they are beyond the scope of this chapter. In the following sections, we will cover four types of amplifiers—erbium-doped fibre amplifiers

(EDFAs), Raman optical amplifiers (ROAs), semiconductor optical amplifiers (SOAs), and amplifiers that are built to amplifier short pulses.

A1.6.3.2 Erbium-doped fibre amplifiers (EDFAs)

Perhaps the most well-known optical amplifier is the EDFA that was first reported in 1987 [12, 40]. Traditionally, the so-called O–E–O signal regeneration was used in optical communication systems, where optoelectronic regenerators are installed between terminals to convert signals from the optical domain to the electrical domain, and then back to the optical domain. Since its invention, EDFA has revolutionized optical communications. Unlike optoelectronic regenerators, this optical amplifier does not require high-speed electronic circuitry and is transparent to data rate and format, which dramatically reduces system cost. EDFAs also provide high gain, high output power, and low noise figure. We will introduce several new important EDFA features and parameters below.

(a) *Energy levels*. The energy levels of the erbium ion and the associated spontaneous lifetime in the fibre glass host are shown in figure A1.6.15. The energy difference between the upper level and the lower level is such that the photons generated are in the 1.5 μm transmission window of an optical fibre. The gain spectra at different inversion levels are shown in figure A1.6.16. Erbium-doped fibre can be pumped by semiconductor lasers at either 980 or 1480 nm. The rapid improvements in semiconductor pump lasers have made the EDFA possible for practical applications. A three-level model can be used for 980 nm pumps while a two-level model usually suffices for 1480 nm pumps [20, 61]. Complete inversion can be achieved with 980 nm pumping but not with 1480 nm pumping [61]. Because of the photon energy difference, the quantum efficiency is higher with 1480 pumping.

(b) *Dynamics*. The spontaneous lifetime of the excited energy level is about 10 ms, which is much slower than the signal bit rates of practical interest. Because of the slow dynamics, an EDFA only experiences the averaged optical power for the practical date rates. As a result of the slow dynamics, inter-symbol distortion and inter-channel crosstalk are negligible even when the EDFAs are working in the saturated region. Besides, since all of the optical signal channels can be amplified simultaneously in one erbium-doped fibre, the EDFA has become a key enabler for the widely used wavelength-division multiplexing (WDM) technology. This is a significant advantage for EDFAs.

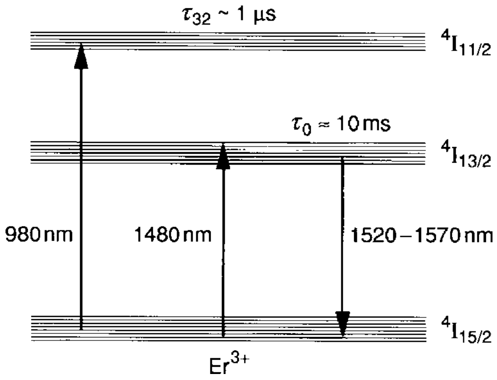


Figure A1.6.15. The energy levels of erbium ion in optical fibre (from [61]).

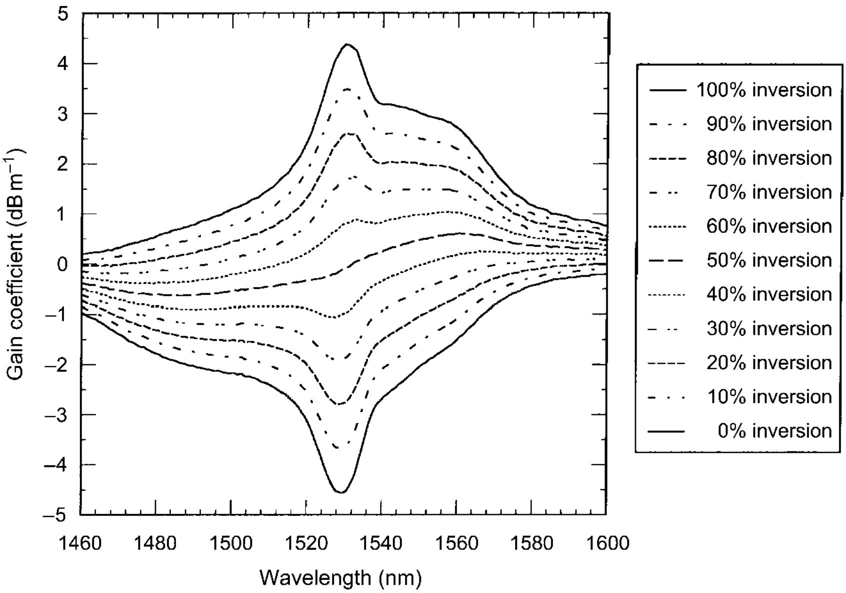


Figure A1.6.16. Gain spectra of erbium-doped fibre (from [61]).

(c) *Bandwidth.* For WDM applications, since uniform gain is desired for all the signal channels, bandwidth is another important parameter for EDFAs. From figure A1.6.16, we can see that the gain is flat somewhere between 1540 and 1560 nm for an inversion level of approximately 50%. Actually, it is this generic flat gain band that was used in initial WDM systems.

Significant progress has been made in amplifier design to achieve excellent performance. Several key techniques are discussed below:

(a) *Multi-stage design.* In order to achieve both low noise figure and high output power, two or more gain stages are usually used where the input stage is kept at a high inversion level and the output stage is kept at a low inversion level [11, 52]. For optical amplifiers with two or more gain stages, the overall noise figure is mainly decided by the high gain input stage and the output power is basically determined by the strongly saturated output stage. The passive components have minimal impact on the noise figure and the output power when they are in the middle stage.

(b) *Gain equalization filter.* A wide bandwidth is desired to accommodate a large number of optical channels. To fully utilize the gain band between 1530 and 1565 nm, gain equalization filters (GEFs) can be used to flatten the gain spectrum. Several technologies have been investigated to fabricate GEFs, including thin film filters, long period gratings, short period gratings, silica waveguide structure, fused fibres, and acoustic filters. Depending on the design, a bandwidth of 35–40 nm can be obtained in the C-band [73]. This kind of amplifier with 35 nm of flat bandwidth was used in the long distance transmission of 32 and 64 channels at 10 Gb s^{-1} [62], and has since become the commonly used bandwidth for C-band communication systems.

(c) *L-band operation.* Because the gain drops sharply on both sides of C-band at the inversion level discussed above, it is not practical to further increase the bandwidth with a GEF. On the other hand, a flat gain region between 1565 and 1615 (L-band) can be obtained at a much lower inversion level [39]. To achieve comparable gain as the C-band, a longer piece of erbium-doped fibre or higher erbium doping would be needed. By combining C-band and L-band, a much wider bandwidth can be realized [63]. Such ultra-wide-band optical amplifiers have made possible high-capacity communication systems at terabits/second or higher [57].

(d) *Large dynamic range.* For commercial systems deployed in the field, the fibre span length and loss varies from location to location. A large dynamic range in the amplifier gain is required for in-line optical amplifiers. Optical attenuators can be used in the middle stage to increase the dynamic range of EDFAs. Such amplifiers can provide flat spectrum and good noise figure when span loss varies.

(e) *Dispersion compensation.* Dispersion compensation is needed for high-speed optical channels and can be done with dispersion compensating fibre. Such compensation is best done on a span-by-span basis. The dispersion compensation modules (DCM) are usually inserted in a middle stage in the EDFAs to optimize the overall optical performance. Typically a middle-stage loss of about 10 dB is reserved for the plug-in of DCM.

In the above discussion, we mainly dealt with the static features of EDFA. In the event of either a network reconfiguration or a failure, the number of WDM signals traversing the amplifiers would change. As a result, the power of the surviving channels would increase or decrease due to the cross-saturation effect in amplifiers. Dropping channels can give rise to surviving channel errors since the power of these channels may surpass the threshold for nonlinear effects such as Brillouin scattering. The addition of channels can cause bit errors by depressing the power of the surviving channels to below the receiver sensitivity. To overcome such error bursts in surviving channels in the network, the signal power transients have to be controlled. Because of the saturation effect, the speed of the gain dynamics in a single EDFA is in general much faster than the spontaneous lifetime of about 10 ms [61]. The time constant of gain dynamics is a function of the saturation caused by the pump power and the signal power. The time constant of gain recovery on single-stage amplifiers was reported to be between 110 and 340 μs [19].

In a recent work, the phenomena of fast power transients in an EDFA chain was reported [76]. Even though the gain dynamics of an individual EDFA is unchanged, the rate of change of the channel power at the end of the system becomes faster for longer amplifier chains. This fast gain dynamics results from the effects of the collective behaviour in a chain of amplifiers. The output of the first EDFA attenuated by the fibre span loss acts as the input to the second EDFA. Since both the output of the first EDFA and the gain of the second EDFA increase with time, the output power of the second amplifier increases at a faster rate. This cascading effect results in faster and faster transients as the number of amplifiers increase in the chain. To prevent performance penalties in a large scale WDM optical network, surviving channel power excursions must be limited to certain values depending on the system margin. Several control schemes have been studied, including pump control [13], link control [56], and laser control [75].

Considerable progress has been made in optical amplifiers in recent years. The bandwidth of amplifiers has increased nearly seven times and flat gain amplifiers with 84 nm bandwidth have been demonstrated. This has been made possible by the addition of the L-band branch. With the advent of these amplifiers, commercial terabits/second lightwave systems will be realizable. Research is underway to develop amplifiers outside the erbium fibre band. Raman amplifiers and semiconductor amplifiers are also potential candidates for amplification across the entire silica fibre transmission band. Progress has also been made in the understanding of gain dynamics of amplifiers. Several control schemes have been successfully demonstrated to mitigate the signal impairments due to fast power transients in a chain of amplifiers and will be implemented in lightwave networks. The terrestrial lightwave systems have been increasing in transmission capacity. To meet the enormous capacity demand the currently available

$400\,\mathrm{Gb\,s^{-1}}$ capacity system with 40 channels will soon be followed by systems having terabits/second and higher capacity on a single fibre.

A1.6.3.3 Raman optical amplifiers

Distributed Raman amplification using optical fibres is an old and yet emerging technology that can supplement the functionality of EDFAs in high-speed ($\geq 10\,\mathrm{Gb\,s^{-1}}$) long-haul transmission systems. Before the advent of the EDFAs, ROAs were used to re-amplify solitons in a re-circulating loop (figure A1.6.17) in order to demonstrate the feasibility of long-haul soliton transmission [42]. Major advantages of using ROAs include low-noise characteristics in distributed mode operation, flexible gain allocation, wide gain bandwidth, and simple construction.

The amount of Raman scattering and its spectrum depend on the material. For silica, which is an amorphous material, the peak Raman gain is at 13.2 THz lower than the pump frequency, which corresponds to a wavelength of about 100 nm longer than the pump at 1450 nm [15, 59]. One can modify the operating gain shape and flatness of the Raman optical amplifier by choosing pump wavelengths and pump powers. For example, one can readily make Raman amplifiers to operate in the S-band (1485–1525 nm) or in the L+ (1604–1640 nm) band. The pump scheme is relatively simple because one only needs to launch the pump light into the fibre via a wavelength-dependent coupler (WDC) or an optical circulator.

The process of Raman scattering can be viewed as the modulation of light (the pump) by molecular vibrations in a material, which are referred to as optical phonons [29]. Upper and lower sidebands appear in the scattered light spectrum because the frequency of the pump is both up-shifted and down-shifted by the optical phonon. At high temperatures, the two sidebands are of equal intensity; however, at room temperature, the lower frequency sideband is favoured. So far we have described the process of

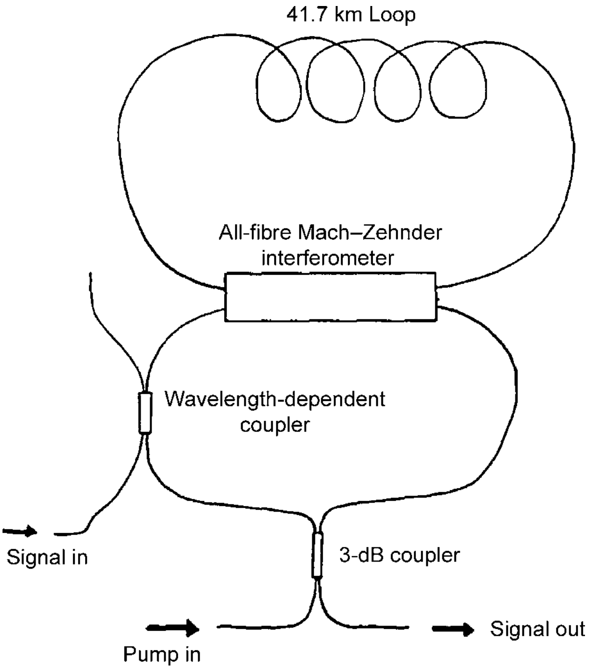

41.7 km Loop
All-fibre Mach–Zehnder interferometer
Wavelength-dependent coupler
Signal in
3-dB coupler
Pump in
Signal out

Figure A1.6.17. The first optical transmission system using Raman optical amplifiers (from [42]).

spontaneous Raman scattering, which was discovered by Sir Raman in 1928 [48]. One can stimulate the process by injecting both the pump and the down-shifted signal into the medium. In this case, the two optical waves will beat together to stimulate the optical phonon, which in turn, will cause more pump to convert into the signal. This is known as stimulated Raman scattering, which was observed 34 years later by Woodbury and Ng [72]. One direct consequence is that the input is amplified coherently and the medium acts like a distributed optical amplifier.

The distributed nature of the Raman amplification provides an advantage to the overall noise performance. As much as 5 dB of improvement in the optical signal-to-noise ratio (OSNR) can be achieved with 10–15 dB of distributed Raman gain. Signal is amplified along the fibre length such that the amount of noise generated in the amplification process is less than that of a lumped amplifier placed at the end of the same fibre span, where signal strength is the weakest. The Raman on–off gain (in decibels), defined as the signal gain with, and without pumps, is proportional to the Raman gain coefficient, the effective length, and the pump intensity. It is also inversely proportional to the fibre effective area (see figure A1.6.18).

Although ROAs hold promise for telecommunication applications, they also have some drawbacks in certain applications compared with those of EDFAs. First, the low efficiency in converting pump power to signal power implies that there is a need for high pump powers. Note that high-power semiconductor pump lasers, which were only available recently, was the main driver for the re-emerging of ROAs. The high-pump-power requirement is generally not desirable for several reasons, such as personnel safety concerns and components reliability concerns. Secondly, the relatively long length of fibre in a ROA allows Rayleigh-related reflection to degrade the performance of a telecommunication system. Typically, the fibre lengths of ROAs are of the order of several kilometres. These Rayleigh-related effects include multiple-path interference (MPI), double-Rayleigh backscattering noise (DRS), and ROA instability, all of which become more severe when pump power is increased [36]. In addition, the fast Raman response time causes time-dependent deleterious effects such as inter-channel cross-talk and the transfer of pump noise to the signal [10]. This is an important constraint for ROAs where the pumps are co-travelling in the same direction as that of the signals. Lastly, similar to EDFAs, ROAs also exhibit transient effects when the input signal power fluctuates [9].

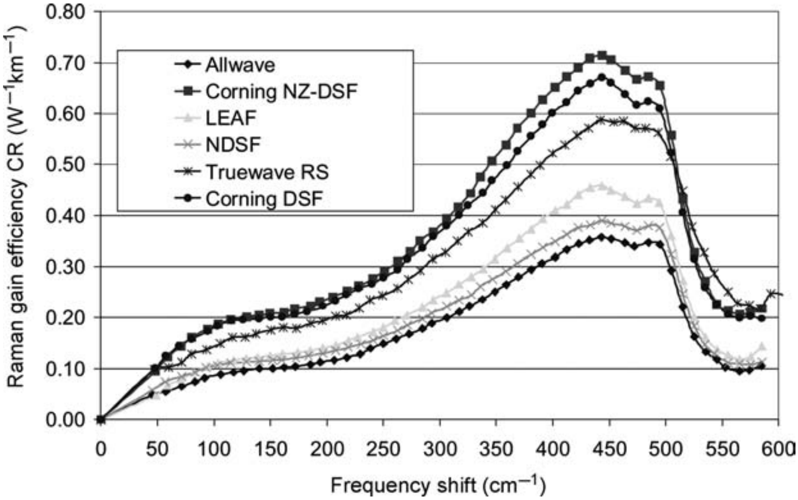


Figure A1.6.18. Raman gain as a function of fibre type (from [15]).

Through careful amplifier and system design, we can overcome some of the problems in ROAs. Nowadays, it is common, especially in laboratories, to combine ROAs and EDFAs to achieve the optimal performance in optical transmission. Several recent $40\,\mathrm{Gb\,s^{-1}}$ transmission experiments rely on distributed Raman gain in order to achieve record-breaking performance. For example, multi-terabits/second transmission experiments over several thousand kilometers were carried out [8, 16, 17, 24].

The quest for higher capacity is, and will still be the focus of transmission research since higher capacity transmission may provide an economic advantage. Increasing the signal-to-noise-ratio using Raman amplifiers is one effective approach. Together with other technologies such as sophisticated coding schemes (forward-error correction, etc), new modulation format, new multiplexing schemes, and higher-quality fibres, the transmission capacity is constantly increasing. Recently, research teams demonstrated $40\,\mathrm{Gb\,s^{-1}}$ transmission using a 100 GHz channel spacing for ultra-long-haul (>3000 km) systems, yielding a high spectral efficiency of $40\,\mathrm{Gb\,s^{-1}}/100\,\mathrm{GHz} = 0.4\,\mathrm{b\,s^{-1}\,Hz^{-1}}$. In the years to come, even more efficient use of available fibre spectrum and higher spectral efficiency are likely to take place. Transmission of multi-terabits/second over long distances will be achieved.

A1.6.3.4 Semiconductor optical amplifiers

SOAs, because of their small physical size, low-cost, and electrical pumping, are ideal amplifiers to be used in low-cost telecommunication systems. To achieve gain in the medium, the semiconductor material is excited to achieve population inversion. There exists two types of SOAs—travelling-wave amplifiers, and Fabry–Perot amplifiers. In the former, the input signal travels from the input port to the output port in the forward direction only, which is made possible by minimizing the reflectivity at the cleaved facets via an antireflection coating. The reflectivity can be reduced further by using angled facets. For Fabry–Perot amplifiers, the input signal is partially reflected at each facet as naturally cleaved facets are used. Because of its resonant nature, the output power of Fabry–Perot SOAs is limited to -10 dBm, while the output power of travelling-wave SOAs can exceed $+13$ dBm.

The gain spectrum of SOAs can be engineered to be as broad as 100 nm. Moreover, through the choice of the material composition (GaAlAs, InGaAlP, InGaAsP, etc), they can operate in the visible or in the near-infrared (1.3–1.6 μm). Since the typical noise figures of SOAs are in the range of 7–9 dB, the use of SOAs has been limited to short-haul (100–200 km) transmission experiments [55, 64].

In a quest to boost output power of a laser, a master-oscillator power-amplifier (MOPA) design is used, where the output of a laser, acting as an oscillator, is amplified subsequently in a booster section. Greater than 1 W of output power is achievable in monolithically integrated MOPAs [41].

SOAs possess undesirable characteristics that must be overcome with clever engineering solutions. For example, the polarization-dependent gain (PDG) of an SOA can be as high as 6 dB. One method to mitigate the polarization effect is to use polarization diversity, where the input optical signal traverses an SOA twice, one in each orthogonal polarization. Another method of growing strained multiple quantum wells in the active region reduces the amount of PDG to below 1 dB [7].

Because the carrier lifetime in SOAs is as short as 500 ps, its fast gain dynamics, together with its relatively low saturation power, induce cross-gain saturation, when the input signal consists of multiple wavelength channels. In other words, if the SOA is used in a WDM system, the instantaneous gain of a given wavelength channel is saturated not only by its channel power, but also by the combined instantaneous power (bit pattern) of the remaining channels. This deleterious effect can be counteracted by (i) operating the SOA in the quasi-linear regime, hence reducing the output power per channel [55], (ii) adding a strong saturating tone to the input signal to act as a reservoir [64], (iii) using polarization multiplexing [58], and (iv) not using intensity modulation on the optical carrier, but using frequency or phase modulation instead [33].

A1.6.3.5 Short-pulse amplification

Although the intensity of a moderate laser beam can be high when it is focused to a spot, on-going work has never stopped in building laser amplifiers to generate higher power laser beams. One limiting factor in amplifying short pulses is the excessive nonlinearity generated when the short pulse passes through the gain medium. Nonlinear effects encountered include (i) self-focusing, which causes the laser beam to collapse to a focal point with catastrophic results; (ii) spatial filamentation of the input beam; (iii) excessive self-phase modulation, where the intensity-dependent phase shift degrades the temporal quality of the output pulse. In other words, in typical high-power short-pulse laser systems, it is the peak intensity, rather than the energy or the fluence, that causes pulse distortion or laser damage. Previously, large laboratories, such as the Lawrence Livermore National Laboratory, construct lasers and amplifiers with large beam diameters as an expensive way to increase the laser power while keeping the intensity below nonlinear effects. CPA overcomes the problem by pre-stretching the input pulse using a dispersive diffraction grating pair and then compressing the output pulse using another pair of diffraction gratings (figure A1.6.19) [60]. Since the optical pulse remains dispersed as it traverses the amplifying medium, the amount of nonlinearity experienced by the optical pulse is minimized. By 1990, this technique has been used to boost the peak power of an incoming laser pulse to over 10 TW (10×10^{12} W) levels (figure A1.6.20). Since the maximum intensity is limited to $10^{12}\,\mathrm{W\,m^{-2}}$ without using CPA, the pulse energy is therefore boosted 10 billion-fold. Although this optical pulse lasts a few tens of femtoseconds only, its instantaneous peak power is equal to the power output from the entire electrical grid in the world. When focused onto a small spot, one can create high-energy pulses with intensities as high as $10^{22}\,\mathrm{W\,m^{-2}}$, thus opening up new and exciting areas in nonlinear optics. For example, electrons are accelerated to relativistic velocities in the so called 'table-top accelerators'. To avoid excessive gain saturation in the amplifier, the repetition rate of the incoming pulse train is often reduced using a pulse-selecting Pockels cell. Incidentally, the concept of CPA is also used in chirped

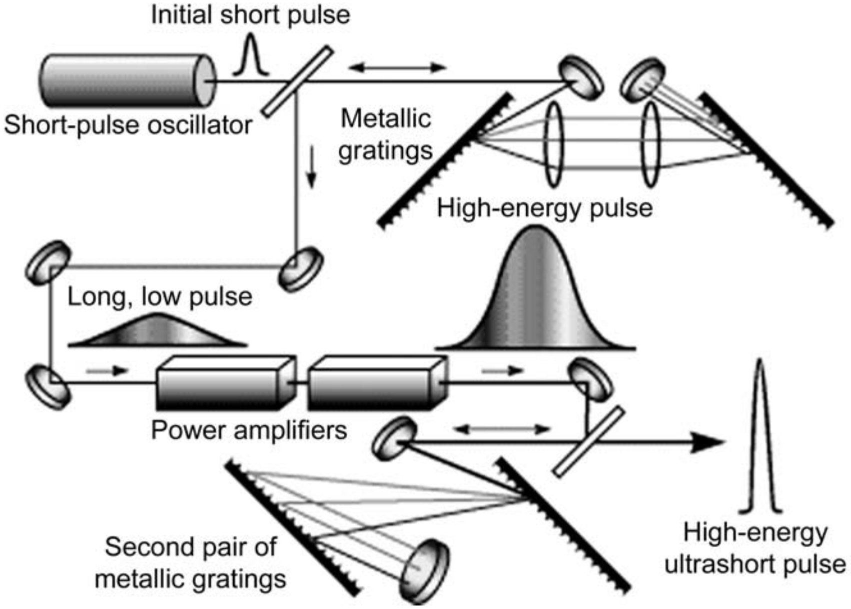


Figure A1.6.19. The concept of chirped pulse amplification (from http://www.llnl.gov/str/Petawatt.html).

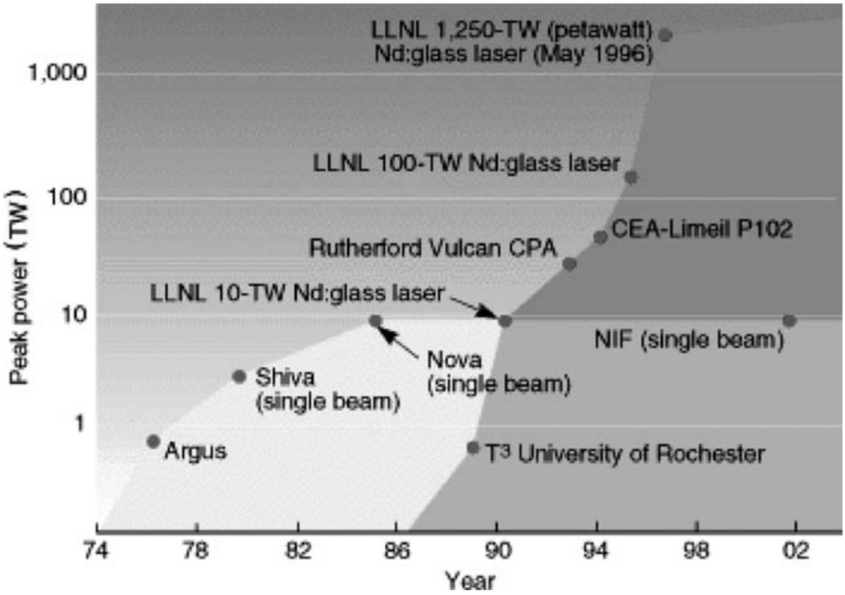


Figure A1.6.20. Milestones in the advancement of laser peak power. The darkest shaded region at upper right indicates that the laser pulses are amplified using chirped pulse amplification (from http://www.llnl.gov/str/Petawatt.html).

radar systems—in order to avoid saturating the power amplifier in the transmitter, the millimetre radar pulse is stretched in time (linearly chirped) by a surface acoustic wave (SAW) acting as a dispersive delay line, before it is amplified and emitted. A pulse-compression filter is then used in the receiver to demodulate and compress the return signal into a shorter pulse, which results in a higher range resolution than radar systems not using chirped pulses.

IMRA
FCPA-2
FCPA-2

Figure A1.6.21. IMRA America, Inc.'s turn-key oscillator-amplifier laser system (from www.imra.com).

Using the latest technological advances in pump laser, rare-earth-doped fibre, control electronics, and packaging, vendors are now offering turn-key table-top systems utilizing CPA. For example, IMRA's oscillator-amplifier laser system can produce high-energy femtosecond pulses with peak powers as much as 10 MW (figure A1.6.21). The diode-pumped system uses CPA in a large-core Yb-doped fibre amplifier. Since optical fibre is an excellent medium to be used as a temperature sensor, the entire system is packaged athermally to achieve good frequency stability and low pulse-timing jitter.

References

[1] Agrawal G P 2002 *Fibre-Optic Communication Systems* 3rd edn (New York: Wiley)

[2] Alferov Z I, Andreev V M, Garbuzov D Z, Zhilyaev Yu V, Morozov E P, Portnoi E L and Trofim V G 1971 Investigation of the influence of the AlAs–GaAs heterostructure parameters on the laser threshold current and the realization of continuous emission at room temperature *Sov. Phys. Semicond.* **4** 1573–1575

[3] Alferov Z 2000 Double heterostructure lasers: early days and future perspectives *IEEE J. Sel. Top. Quantum Electron.* **6** 832–840

[4] Boyd G D and Kogelnik H 1962 Generalized confocal resonator theory *Bell System Tech. J.* **41** 1347

[5] Bennett W R 1962 Gaseous optical masers *Appl. Opt. Suppl.* **1** *Optical Masers* 24

[6] Bridges W B 1964 Laser oscillation in singly ionized argon in the visible spectrum *Appl. Phys. Lett.* **4** 128

[7] Cole S, Cooper D M, Devlin W J, Ellis A D, Elton D J, Isaak J J, Sherlock G, Spurdens P C and Stallard W A 1989 *Electron. Lett.* **25** 314–315

[8] Charlet G *et al* 2002 6.4 Tb/s (159 × 42.7 Gb/s) capacity over 21 × 100 km using bandwidth-limited phase-shaped binary transmission *Proc. European Conference on Optical Communication* (Copenhagen, Denmark, 2002) paper PD4.1

[9] Chen C-J and Wong W S 2001 Transient effects in saturated Raman amplifiers *Electron. Lett.* **37** 371

[10] Chraplyvy A R and Henry P S 1983 Performance degradation due to stimulated Raman scattering in wavelength-division-multiplexed optical fiber systems *Electron. Lett.* **19** 641–643

[11] Delavaux J-M P and Nagel J A 1995 Multi-stage erbium-doped fiber amplifier designs *J. Lightwave Technol.* **13** 703–720

[12] Desurvire E, Simpson J R and Becker P C 1987 High-gain erbium-doped traveling-wave fiber amplifier *Opt. Lett.* **12** 888–890

[13] Desurvire E, Zirngibl M, Presby H M and DiGiovanni D 1991 Dynamic gain compensation in saturated erbium-doped fiber amplifiers *IEEE Photon. Technol. Lett.* **3** 453–455

[14] Einstein A 1917 On the quantum theory of radiation *Physikalische Zeitschrift* **18** 121

[15] Fludger C, Maroney A, Jolley N and Mears R 2000 An analysis of improvements in OSNR from distributed Raman amplifiers using modern transmission fibers *Proc. Optical Fibre Communications Conference* (Baltimore, USA, 2000) paper FF2

[16] Foursa D G *et al* 2002 2.56 Tb/s (256 × 10 Gb/s) transmission over 11000 km using hybrid Raman/EDFAs with 80 nm of continuous bandwidth *Proc. Optical Fiber Communications Conference* (Anaheim, USA, 2002) postdeadline paper FC3

[17] Gnauck A H *et al* 2002 2.5 Tb/s (64 × 42.7 Gb/s) transmission over 40 × 100 km NZDSF using RZ-DPSK format and all-Raman-amplified spans *Proc. Optical Fiber Communications Conference* (Anaheim, USA, 2002) postdeadline paper FC2

[18] Geusic J E, Marcos H M and Van Uitert L G 1964 Laser oscillations in Nd-doped yttrium aluminum, yttrium gallium and gadolinium garnets *Appl. Phys. Lett.* **4** 182

[19] Giles C R, Desurvire E and Simpson J R 1989 Transient gain and cross-talk in erbium-doped fiber amplifier *Opt. Lett.* **14** 880–882

[20] Giles C R and Desurvire E 1991 Modeling erbium-doped fiber amplifiers *J. Lightwave Technol.* **9** 271–283

[21] Gordon E I, Labuda E F and Bridges W B 1964 Continuous visible laser action in singly ionized Argon, Krypton and Xenon *Appl. Phys. Lett.* **4** 178

[22] Gordon J P, Zeiger H J and Townes C H 1954 Molecular microwave oscillator and new hyperfine structure in the microwave spectrum of NH_3 *Phy. Rev.* **94** 282–284

[23] Gordon J P, Zeiger H J and Townes C H 1955 The Maser—new type of microwave amplifier, frequency standard, and spectrometer *Phy. Rev.* **95** 1264–1274

[24] Grosz D F *et al* 2002 5.12 Tb/s (128 × 42.7 Gb/s) transmission with 0.8 bit/s/Hz spectral efficiency over 1280 km of standard single-mode fiber using all-Raman amplification and strong signal filtering *Proc. European Conf. Optical Communication* (Copenhagen, Denmark, 2002) paper PD4.3

[25] Hall R N, Fenner G E, Kingsley J D, Soltys T J and Carlson R O 1962 *Phys. Rev. Lett.* **9** 366

[26] Haus H A 1975 A theory of forced modelocking *IEEE J. Quantum Electron.* **11** 323–330

[27] Haus H A 1984 *Waves and Fields in Optoelectronics* (Englewood Cliffs: Prentice-Hall)

[28] Hayashi I, Panish M B, Foy W and Sumski S 1970 *Appl. Phys. Lett.* **17** 109

[29] Hellwarth R W 1963 *Phys. Rev.* **130** 1850

[30] Holonyak N Jr and Bevacqua S F 1962 *Appl. Phys. Lett.* **1** 82

[31] Javan A, Bennett W R Jr and Herriott D R 1961 Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture *Phys. Rev. Lett.* **6** 106

[32] Kane D J and Trebino R 1993 Characterization of arbitrary femtosecond pulses using frequency-resolved optical gating *IEEE J. Quantum Electron.* **29** 571–579

[33] Kim H K, Chandrasekhar S, Srivastava A, Burrus C A and Buhl L 2001 10 Gbit/s based WDM signal transmission over 500 km of NZDSF using semiconductor optical amplifier as the in-line amplifier *Electron. Lett.* **37** 185–187

[34] Koechner W 1976 *Solid State Engineering* (New York: Springer)

[35] Kushida T, Marcos H M and Geusic J E 1968 Laser transition cross section and fluorescence branching ratio for Nd^{3+} in yttrium aluminum garnet *Phys. Rev.* **167** 289

[36] Lewis S A E, Chernikov S V and Taylor J R 2000 Characterization of double Rayleigh scatter noise in Raman amplifiers *IEEE Photon. Technol. Lett.* **12** 528

[37] Maiman T H 1960 Stimulated optical radiation in ruby *Nature* **187** 493

[38] Maiman T H 1960 Optical and microwave-optical experiments in ruby *Phys. Rev. Lett.* **4** 564

[39] Massicott J F, Armitage J R, Wyatt R, Ainslie B J and Craig-Ryan S P 1990 High gain, broadband, 1.6 μm Er^{3+} doped silica fiber amplifier *Electron. Lett.* **20** 1645–1646

[40] Mears R J, Reekie L, Jauncey I M and Payne D N 1987 Low-noise erbium-doped fiber amplifier operating at 1.54 μm *Electron. Lett.* **23** 1026–1028

[41] Mehuys D, Parke R, Waarts R G, Welch D F, Hardy A and Streifer W 1991 Characteristics of multistage monolithically integrated master oscillator power amplifiers *IEEE J. Quantum Electron.* **27** 1574–1581

[42] Mollenauer L F and Smith K 1988 Demonstration of soliton transmission over more than 4000 km in fiber with loss periodically compensated by Raman gain *Opt. Lett.* **13** 675–677

[43] Moulton P F 1986 Spectroscopic and laser characteristics of Ti:Al_2O_3 *J. Opt. Soc. Amer. B* **3** 125–133

[44] Nathan M I, Dumke W P, Burns G, Dill F H Jr and Lasher G 1962 *Appl. Phys. Lett.* **1** 62

[45] Patel C K N 1964 Introduction of CO_2 optical maser experiments *Phys. Rev.* **136** A1187

[46] Patel C K N 1968 High power carbon dioxide lasers *Sci. Am.* **219** 22–23

[47] Quist T M, Rediker R H, Keyes R J, Krag W E, Lax B, McWhorter A L and Zeiger H J 1962 *Appl. Phys. Lett.* **1** 91

[48] Raman C V 1928 *Indian J. Phys.* **2** 387

[49] Saleh B E A and Teich M C 1991 *Fundamental of Photonics* (New York: Wiley)

[50] Schawlow A L and Townes C H 1958 Infrared and optical masers *Phys. Rev.* **112** 1940–1949

[51] Siegman A E 1986 *Lasers* (Mill Valley: University Science Books)

[52] Smart R G, Zyskind J L and DiGiovanni D J 1994 Two-stage erbium-doped fiber amplifiers suitable for use in long-haul soliton systems *Electron. Lett.* **30** 50–52

[53] Smith P W 1970 Mode-locking of lasers *Proc. IEEE* **58** 1342–1359

[54] Snitzer E and Young C G 1968 Glass lasers *Lasers* vol 2, ed A K Levine (New York: Dekker) p 191

[55] Spiekman L H, Wiesenfeld J M, Gnauck A H, Garret L D, van den Hoven G N, van Dongen T, Sander-Jochem M J H and Binsma J J M 2000 8 × 10 Gb/s DWDM transmission over 240 km of standard fiber using a cascade of semiconductor optical amplifiers *IEEE Photon Technol. Lett.* **12** 1082–1084

[56] Srivastava A K *et al* 1997 Fast-link control protection of surviving channels in multiwavelength optical networks *IEEE Photon. Technol. Lett.* **9** 1667–1669

[57] Srivastava A K *et al* 1998 1 Tb/s transmission of 100 WDM 10 Gb/s channels over 400 km of TrueWave™ fiber *OFC '98 Technical Digest* (San Jose, USA, 1998) postdeadline paper PD-10

[58] Srivastava A K, Banerjee S, Eichenbaum B R, Wolf C, Sun Y, Sulhoff J W and Chraplyvy A R 2000 A polarization multiplexing technique to mitigate WDM crosstalk in SOAs *IEEE Photon. Technol. Lett.* **12** 1415–1416

[59] Stolen R H 1980 *Proc. IEEE* **68** 1232

[60] Strickland D and Mourou G 1985 Compression of amplified chirped optical pulses *Opt. Commun.* **55** 447

[61] Sun Y, Zyskind J L and Srivastava A K 1997 Average saturation level, modeling, and physics of erbium-doped fiber amplifiers *IEEE J. Sel. Top. Quantum Electron.* **3** 991–1007

[62] Sun Y *et al* 1997 Transmission of 32-WDM 10-Gb/s channels over 640 km using broad-band, gain-flattened erbium-doped silica fiber amplifiers *IEEE Photon Technol. Lett.* **9** 1652–1654

[63] Sun Y *et al* 1997 Ultra-wide-band erbium-doped silica fiber amplifier with 80 nm of bandwidth *Optical Amplifiers and their Applications* (Victoria, Canada, 1997) postdeadline paper PD-2

[64] Sun Y, Srivastava A K, Banerjee S, Sulhoff J W, Pan R, Kantor K, Jopson R M and Chraplyvy A R 1999 Error-free transmission of 32 × 2.5 Gb/s DWDM channels over 125 km using cascaded in-line semiconductor optical amplifiers *Electron. Lett.* **35** 1863–1865

[65] Tolman R C 1924 Duration of molecules in upper quantum states *Rev. Mod. Phys.* **23** 693–709

[66] Townes C H 1964 Production of coherent radiation by atoms and molecules *Nobel Lectures Physics 1963–1970* (Amsterdam: Elsevier) pp 58–85

[67] Verdeyen J T 1989 *Laser Electronics* 2nd edn (Englewood Cliffs: Prentice-Hall)

[68] Vuylsteke A A 1963 Theory of laser regeneration switching *J. Appl. Phys.* **34** 1615

[69] Wagner W G and Lengyel B A 1963 Evolution of the giant pulse in a laser *J. Appl. Phys.* **34** 2040

[70] Walmsley I and Trebino R 1996 Measuring fast pulses with slow detectors *Opt. Photon. News* **7** 23

[71] Weber H P 1967 Method for pulsewidth measurement of ultrashort light pulses generated by phase-locked lasers using nonlinear optics *J. Appl. Phys.* **38** 2231–2234

[72] Woodbury E J and Ng W K 1962 *Proc. IRE* **50** 2347

[73] Wysocki P F, Judkins J B, Espindola R P, Andrejco M and Vengsarkar A M 1997 Broad-band erbium-doped fiber amplifier flattened beyond 40 nm using long-period grating filter *IEEE Photon. Technol. Lett.* **9** 1343–1345

[74] Yariv A 1985 *Optical Electronics* 3rd edn (New York: Holt, Rinehart, and Winston)

[75] Zirngibl M 1991 Gain control in erbium-doped fiber amplifiers by an all-optical feedback loop *Electron. Lett.* **27** 560–561

[76] Zyskind J L, Sun Y, Srivastava A K, Sulhoff J W, Lucero A L, Wolf C and Tkach R W 1996 Fast power transients in optically amplified multiwavelength optical networks *Proc. Optical Fiber Communications Conference* (San Jose, USA, 1996) postdeadline paper PD-31

A2.1
Advanced optics

Alan Rogers

A2.1.1 Introduction

The practice of optoelectronics at the level of device and system design necessitates a familiarity with some quite advanced, and sometimes quite subtle, optical physics. This chapter will deal, at a fairly fundamental level, with those topics which the author considers most important for an appreciation of present-day optoelectronics.

The topics chosen naturally are concerned with the properties of light and its control by waveguides, and with those properties of solid materials relevant to their interactions with light radiation.

Many of the basic ideas which are needed to understand these processes have already been introduced in earlier chapters; in this chapter, however, we shall need to extend and refine the coverage to the point where we can understand the structure and operation of specific optoelectronic devices. These devices depend, in most cases, on the physics of optical radiation and of the solid state, in a fairly detailed way.

It is thus necessary to gain a more detailed familiarity with this physics, since it lies at the heart of the subject, and without this knowledge it would be impossible either to understand properly existing optoelectronics or to progress to new devices and new systems beyond our present-day thinking.

We shall begin by looking at the physics of radiation.

A2.1.2 The physics of radiation

A2.1.2.1 Black-body radiation

All matter, provided that it has a temperature other than absolute zero, emits radiation. This is a consequence of the fact that a temperature above absolute zero implies that the atoms or molecules are in motion, and are thus colliding with each other constantly. These collisions transfer kinetic energy of motion, but also sometimes excite the atomic system to a higher state, from which it may relax by emitting a photon. This is a consequence of a very basic principle in physics, the ‘law of equipartition of energy’, which states that the energy of a system, in equilibrium, will be distributed equally among all possible degrees of freedom: the kinetic energy of a material is one such degree, excited states represent another. By assigning a temperature to a body, we require it to be in thermal equilibrium with its surroundings (i.e. there is no net heat gain from, or loss to, the surroundings over time), so equipartition must apply.

The first question which naturally arises now is: how much radiation is emitted by a body at a given temperature? And the second (perhaps not quite as naturally!) is: what is the distribution of this emitted radiation over the wavelength spectrum?

In answering these questions, we shall explore ideas which are valuable for a whole range of topics in optoelectronics, and more general physics, so it is worth while taking some time over them.

Classical thermodynamics assumed that atoms emitted light as a result of radiation by electrons which were oscillating at natural resonant frequencies within the atoms. It further assumed (it had no reason to assume otherwise) that these oscillations could occur with any amount of energy, depending on the strength of the stimulus.

The other piece of information which the classical thermodynamicists needed before they could proceed was the Boltzmann factor. This tells us the ratio of the number of atoms which have energy E_1 compared to those with energy E_2 in a system in equilibrium at absolute temperature T, and takes the form [1]:

$$\frac{N_1}{N_2} = \exp\left(-\frac{E_1 - E_2}{kT}\right)$$

where k is Boltzmann's constant, with value $1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$. This factor had already been derived, using classical statistical thermodynamics, by means of an exquisite argument (which we do not have space to develop, but see any text on statistical mechanics).

Let us now derive the classical result for the radiation emitted by a 'perfect' body, i.e. a body capable of emitting or absorbing radiation of any wavelength and thus containing oscillators (atoms or molecules) capable of oscillating at any frequency. Such a body is called a 'black' body since it absorbs, rather than reflects, all light falling upon it, and therefore looks 'black' (until it emits!). Such a body is a valuable idealization, since we can categorize 'real' bodies according to how closely they approximate to it.

Let us suppose that within this black body the oscillators can have any energy and (for reasons which will become clear later) these energies will be described as a set distributed as:

$$0, \mathrm{d}E, 2\mathrm{d}E, \cdots, n\mathrm{d}E, \cdots$$

where dE is infinitesimally small, so that the distribution is continuous. The ratios of numbers of oscillators with each of these energy bands will comprise the set (according to the Boltzmann factor):

$$1 : \exp\left(-\frac{\mathrm{d}E}{kT}\right) : \exp\left(-\frac{2\mathrm{d}E}{kT}\right) : \cdots : \exp\left(-\frac{n\mathrm{d}E}{kT}\right) : \cdots.$$

Thus if there are N oscillators with zero energy, the total number of oscillators will be:

$$N_\mathrm{T} = N\left[1 + \exp\left(-\frac{\mathrm{d}E}{kT}\right) + \exp\left(-\frac{2\mathrm{d}E}{kT}\right) + \cdots + \exp\left(-\frac{n\mathrm{d}E}{kT}\right) + \cdots\right].$$

This is a geometrical progression which is easily summed to give:

$$N_\mathrm{T} = \frac{N}{1 - \exp\left(-\dfrac{\mathrm{d}E}{kT}\right)}. \tag{A2.1.1}$$

Also, the total energy of all the oscillators is given by multiplying each term by its energy allocation:

$$E_\mathrm{T} = N\left[\mathrm{d}E\exp\left(-\frac{\mathrm{d}E}{kT}\right) + 2\mathrm{d}E\exp\left(-\frac{2\mathrm{d}E}{kT}\right) + \cdots\right.$$
$$\left. + n\mathrm{d}E\exp\left(-\frac{n\mathrm{d}E}{kT}\right) + \cdots\right]$$

giving, on summation,

$$E_{\mathrm{T}} = \frac{N\mathrm{d}E}{\exp\left(\frac{\mathrm{d}E}{kT}\right)\left[1 - \exp\left(-\frac{\mathrm{d}E}{kT}\right)\right]^2}. \tag{A2.1.2}$$

On dividing equation (A2.1.2) by equation (A2.1.1) we obtain the mean energy per oscillator as:

$$\bar{E} = \frac{\mathrm{d}E}{\exp\left(\frac{\mathrm{d}E}{kT}\right) - 1}. \tag{A2.1.3}$$

We may now let $\mathrm{d}E \rightarrow 0$ to discover the physical value for this mean energy, whereupon we find (expanding the exponential in the denominator):

$$\bar{E} = \lim_{\mathrm{d}E\rightarrow 0} \frac{\mathrm{d}E}{1 + \frac{\mathrm{d}E}{kT} + \frac{1}{2}\left(\frac{\mathrm{d}E}{kT}\right)^2 + \cdots - 1}$$

or

$$\bar{E} = kT. \tag{A2.1.4}$$

(This is entirely in accordance with other 'equipartitional' approaches which allow $kT/2$ of energy per degree of freedom. In our case there are two degrees of freedom per oscillator, one for kinetic energy, the other for potential energy, giving kT in all.)

The final piece of information we need is the number of independent oscillations which can occur within a given volume of material. Clearly, the fact that the volume is finite means that there are boundaries and these impose boundary conditions on the oscillations, just as a string stretched between two fixed points is bounded by the fact that any oscillation of the string must have zero amplitude at the points of fixation.

Let us simplify things by taking the volume to be a cube of side l (figure A2.1.1(*a*)): suppose now that oscillations occur within the cube and that the velocity with which these propagate is c. The walls of the cube impose a zero-amplitude boundary condition for these oscillations, so the resonant oscillations can only occur parallel with the sides of the cube with frequencies $nc/2l$, n being a positive integer and $c/2l$ the fundamental.

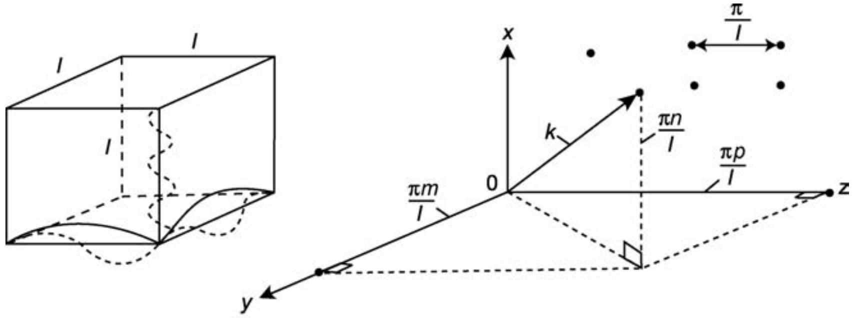
Now waves can, of course, travel in many directions and we can best represent any given wave by its wave vector $\boldsymbol{k}$, which has the same direction of the wave and an amplitude $2\pi/\lambda$, where λ is the wavelength. We may now write the frequencies of the waves travelling parallel to the sides of the cube as:

$$f_n = \frac{nc}{2l}$$

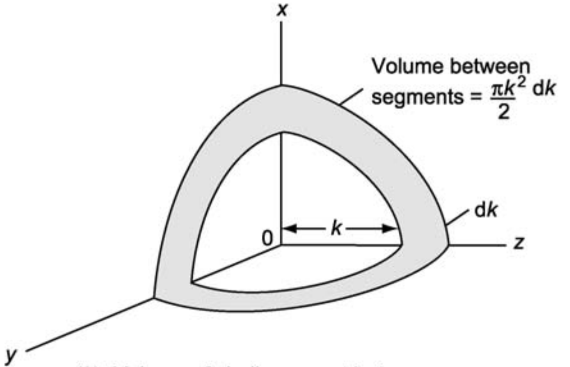
and the wavenumbers as:

$$k_n = \frac{2\pi}{c} f_n = \frac{\pi n}{l}. \tag{A2.1.5}$$

Let us now take axes Ox, Oy, Oz parallel with the sides of the cube and plot in three dimensions a lattice of points corresponding to all the k_n in equation (A2.1.5). It is easily seen that *any* oscillation for the cube can be represented by its wave vector from the origin of axes to one of the points we have plotted. The plot is often called 'k-space', for obvious reasons.



(*a*) Waves in a bounded cube, and '*k*-space'



(*b*) Volume of shell segment in *k*-space

Figure A2.1.1. *k*-space diagram.

We now ask the question: how many oscillations can the cube support with wavelengths between k and $k + \mathrm{d}k$? This, we can see, will correspond to the number of points on our plot which lie in the volume between spheres of radius k and $k + \mathrm{d}k$, respectively. This volume in k-space is, $4\pi k^2 \mathrm{d}k$ (figure A2.1.1(*b*)). However, since only positive values of n are valid we only need that octant of the spherical shell where all the axes are positive, which is one-eighth of the total, i.e. $\pi k^2 \mathrm{d}k/2$. To find the number of points in this volume we divide by one elementary volume in our lattice, defined by the interval between points, i.e. a cube of side π/l, volume π^3/l^3.

Hence the number of oscillations between k and $k + \mathrm{d}k$ is

$$N_0' = \frac{\frac{1}{2}\pi k^2 \mathrm{d}k}{\pi^3/l^3} = \frac{k^2 l^3 \mathrm{d}k}{2\pi^2}$$

and if we allow two orthogonal linear polarizations per oscillation (any electromagnetic wave can always be resolved into two such components) this becomes:

$$N_0 = \frac{k^2 l^3 \mathrm{d}k}{\pi^2}.$$

Since the volume of our original cube is l^3 we can express this in the form of a number of oscillations per unit volume:

$$N_{\mathrm{v}} = \frac{k^2 \mathrm{d}k}{\pi^2}. \tag{A2.1.6a}$$

It is now more convenient to write N_{v} in terms of frequency (since frequency is more directly related with energy). We have

$$k = \frac{2\pi}{\lambda} = \frac{2\pi f}{c}.$$

Hence

$$N_{\mathrm{v}} = \frac{8\pi f^2}{c^3} \mathrm{d}f. \tag{A2.1.6b}$$

This is an important result in itself and appears in many aspects of laser theory; it should be noted carefully.

We may use it immediately for our present purposes to derive the classical result for the energy spectrum of a black-body radiator. From equation (A2.1.4) we saw that each oscillation has mean energy kT. Hence the energy density (i.e. energy per unit volume) lying between frequencies f and $f + \mathrm{d}f$ is given by:

$$\rho_{\mathrm{f}} \mathrm{d}f = \frac{8\pi f^2}{c^3} kT \mathrm{d}f. \tag{A2.1.7}$$

This is the classical result, the so-called Rayleigh–Jeans equation. *It is wrong!* It has to be. We can see this immediately by calculating the total energy density emitted over all wavelengths of the spectrum:

$$\rho_{\mathrm{T}} = \int_0^\infty \frac{8\pi f^2}{c^3} kT \mathrm{d}f$$

$$= \frac{8\pi kT}{c^3} \left[\frac{f^3}{3} \right]_0^\infty \rightarrow \infty!$$

The answer is thus that an infinite amount of total energy is emitted by any black body; this is, of course, quite impossible. This result caused much head scratching amongst classical physicists around the turn of the century. When the spectrum of a black body (or as close to one as could be realized in practice) was measured, the shape was as shown in figure A2.1.2. The agreement with the Rayleigh–Jeans expression (equation (A2.1.7)) was good at low frequencies, but the two diverged wildly at the higher frequencies: the problem was thus dubbed the 'ultra-violet (i.e. high frequency) catastrophe'.

A2.1.2.2 The quantum result

Max Planck, in 1900, saw a very simple way to avoid the above problem. He suggested that an oscillator could not possess *any* value of energy but only a value which was an integer times a certain minimum value. If this latter was ε, then the only possible values for the energy of the oscillation were ε, 2ε, $3\varepsilon \cdots n\varepsilon$. This changed things completely, as we shall now show.

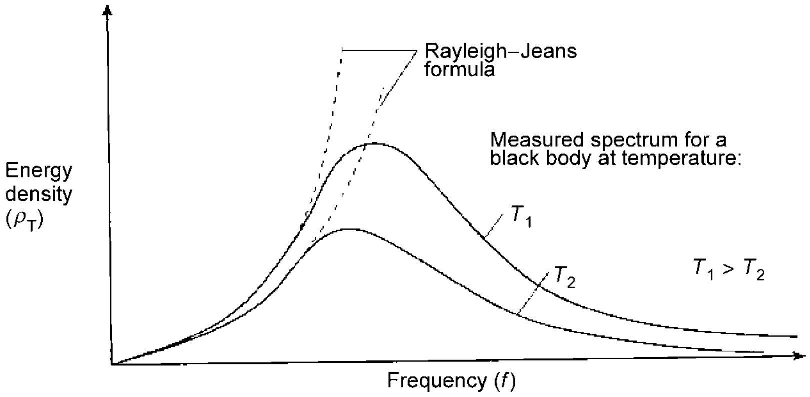


Figure A2.1.2. The black-body spectrum.

We can now, very conveniently, return to equation (A2.1.3) for the mean energy of an oscillator:

$$\bar{E} = \frac{\mathrm{d}E}{\exp\left(\dfrac{\mathrm{d}E}{kT}\right) - 1}$$

dE can now be identified with ε, and now it does *not* tend to zero, but always remains nonzero.

Now ε is the minimum energy of the oscillator which emits radiation and thus can be identified, in turn, with the quantity $h\nu$, where ν is the lowest frequency of radiation it emits, remembering that the oscillator can have any of the energies $nh\nu$, where n is any positive integer. (We now use ν for frequency, rather than f, to remind ourselves that we are dealing with quantum phenomena rather than continuous events, h is the quantum constant (Planck's constant) with value 6.626×10^{-34} J s.)

Thus we have, in the quantum case:

$$\bar{E} = \frac{h\nu}{\exp\left(\dfrac{h\nu}{kT}\right) - 1}$$

rather than kT, and the energy density lying between ν and $\nu + \mathrm{d}\nu$ now will be, using equation (A2.1.7):

$$\rho_\nu \mathrm{d}\nu = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{\exp\left(\dfrac{h\nu}{kT}\right) - 1} \mathrm{d}\nu. \tag{A2.1.8}$$

This is the celebrated Planck radiation formula, and it solves all our problems, for it agrees with the experimental spectrum (figure A2.1.2).

If integrated over all frequencies it remains finite, and gives the result:

$$E_\mathrm{T} = \frac{2\pi^5 k^4}{15 c^2 h^3} T^4. \tag{A2.1.9}$$

Equation (A2.1.9) represents the Stefan–Boltzmann law for the total energy emitted by a black body; classical thermodynamics was able to show that this quantity should be proportional to the fourth power of the absolute temperature, but was unable to predict the value of the constant of proportionality; quantum physics has provided the answer to this.

Similarly, classical thermodynamics was able to prove Wien's displacement law, which states that the value of the wavelength associated with the energy maximum in the spectrum (figure A2.1.2) is inversely proportional to the absolute temperature, i.e.

$$\lambda_{\mathrm{m}} = \frac{\Omega}{T}$$

but was unable to determine the value of the constant Ω. By differentiating equation (A2.1.8) we easily find that:

$$\Omega = \frac{ch}{4.9651k}.$$

The above results had a profound effect. Although Planck at first felt that his quantum hypothesis was no more than a mathematical trick to avoid the ultra-violet catastrophe, it soon became clear that it was fundamentally how the universe did, in fact, behave: quantum theory was born.

A2.1.2.3 'Black-body' sources

The concept of a black body is that of a body which emits and absorbs all frequencies of radiation. We know now that the quantum theory requires us to limit the frequency to multiples of a certain fundamental frequency, but, in practice, owing to the particular molecular structure of any given body, the quantum (and classical) 'black body' remains an idealization, and real bodies, when hot, will not yield a spectrum in strict accordance with Planck's radiation law but only an approximation to it (sometimes a very close approximation, however).

Nevertheless, we can very conveniently measure the temperature of a radiating body by measuring the wavelength at which the spectrum peaks, using Wien's law, or, if the peak is not at a convenient (for our detector) position in the spectrum, by measuring the total energy emitted (using a bolometer) and applying the Stefan–Boltzmann law. Very often we require a source which emits over a broad range of frequencies, and a convenient way to obtain this is to create a discharge in a gas. An electrical discharge creates a large number of free, energetic electrons which cause a large range of atomic excitations, thus giving rise to radiation over a broad frequency range. Intensities can be quite high, so that the experimenter or designer can then pick out those frequencies that are needed, with frequency-selective optical components such as prisms or diffraction gratings.

However, the importance of the idealization known as a black body lies primarily in the fact that it allows an insight into the fundamental nature of electromagnetic radiation and the quantum laws which it obeys. This is crucial to our understanding of its role in optoelectronics, and especially to our understanding of laser radiation, which is the next topic for consideration.

A2.1.2.4 The theory of laser action

The rate equations and the gain mechanism

The elements of laser action were introduced in chapter A.1. Lasers are extremely important in optoelectronics, as has been stressed, and it is necessary now to deal with laser action in more quantitative detail.

Let us consider two energy levels of an atomic system E_1 and E_2, with $E_2 > E_1$ (figure A2.1.3).

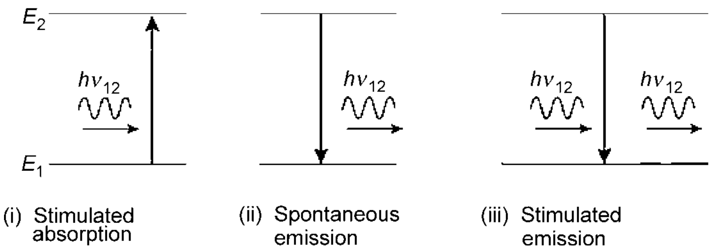

E_2
E_1
$h\nu_{12}$
$h\nu_{12}$
$h\nu_{12}$
$h\nu_{12}$
(i) Stimulated absorption
(ii) Spontaneous emission
(iii) Stimulated emission

Figure A2.1.3. Two-level photon transitions.

We know that the system can be raised from E_1 to E_2 by absorption of a photon with frequency ν_{12}, where:

$$h\nu_{12} = E_2 - E_1$$

and we also know that the system, after having been excited, will eventually, spontaneously, revert to its ground state E_1 by emitting a photon of energy $h\nu_{12}$.

However, the excited state E_2 can also be stimulated to decay to the state E_1 by the action of another photon of energy $h\nu_{12}$. This process is called stimulated emission. Thus now we are considering three distinct processes:

(a) stimulated absorption ($E_1 \rightarrow E_2$);

(b) spontaneous emission ($E_2 \rightarrow E_1$);

(c) stimulated emission ($E_2 \rightarrow E_1$).

There can be no spontaneous 'absorption' since this would violate the law of conservation of energy.

In order to calculate the relationships between atoms and radiation in equilibrium (i.e. black-body radiation), Einstein used a very simple argument: consider the atoms to be in equilibrium with each other and with the radiation in a closed system. The rate (per unit volume) at which atoms are raised to the upper state is proportional to the density of photons, ρ_ν, at energy $h\nu_{12}$ and to the density of atoms N_1 (number per unit volume) in state E_1, i.e.

$$R_{12} = N_1 \rho_\nu B_{12} \quad \text{(stimulated absorption)}$$

where B_{12} is a constant.

Similarly, the rate at which atoms in state E_2 are stimulated to return to state E_1 is given by:

$$R_{21} = N_2 \rho_\nu B_{21} \quad \text{(stimulated emission)}$$

where N_2 is the density of atoms in state E_2. Now spontaneous emission from state E_2 to E_1 occurs after a characteristic delay determined by the detailed atomic characteristics, and is governed by quantum rules. Its rate therefore is proportional to N_2, the constant of proportionality, comprising, essentially, the reciprocal of the decay time. Thus we have

$$S_{21} = N_2 A_{21} \quad \text{(spontaneous emission).}$$

The constants A_{21}, B_{12}, B_{21} are called the Einstein coefficients.

Clearly, in equilibrium, we must have:

$$N_1\rho_\nu B_{12} = N_2\rho_\nu B_{21} + N_2 A_{12} \tag{A2.1.10}$$

since the total upward and downward transition rates must be equal.

Hence from equation (A2.1.10):

$$\rho_\nu = \frac{(A_{21}/B_{21})}{(B_{12}N_1/B_{21}N_2) - 1}.$$

But we know from the Boltzmann relation that:

$$\frac{N_1}{N_2} = \exp\left(-\frac{E_1 - E_2}{kT}\right)$$

and also that $E_2 - E_1 = h\nu_{12}$.

Hence, generalizing from ν_{12} to ν:

$$\rho_\nu = \frac{(A_{21}/B_{21})}{(B_{12}/B_{21})\exp\left(\dfrac{h\nu}{kT}\right) - 1}. \tag{A2.1.11}$$

Now it was shown in section A2.1.2.2 that for equilibrium (black-body) radiation (equation A2.1.8):

$$\rho_\nu = \frac{8\pi h\nu^3}{c^3}\frac{1}{\exp\left(\dfrac{h\nu}{kT}\right) - 1}.$$

Hence it follows, by comparing this with (A2.1.11):

$$B_{12} = B_{21} \tag{A2.1.12a}$$

$$A_{21} = B_{21}\frac{8\pi h\nu^3}{c^3}. \tag{A2.1.12b}$$

Relations (A2.1.12*a*) and (A2.1.12*b*) are known as the Einstein relations, and are very important determinants in the relationships between atoms and radiation. For example, it is clear that, under these conditions, the ratio of stimulated to spontaneous emission from E_2 to E_1 is given by:

$$S = \frac{R_{21}}{S_{21}} = \frac{\rho_\nu N_2 B_{21}}{N_2 A_{21}} = \frac{\rho_\nu c^3}{8\pi h\nu^3}$$

and using the expression for ρ_ν from equation (A2.1.8):

$$S = \frac{1}{\exp\left(\dfrac{h\nu}{kT}\right) - 1}.$$

If, for example, we consider the specific case of the He–Ne discharge at a temperature of 370 K with $\lambda = 632.8$ nm ($\nu = 4.74 \times 10^{14}$ Hz) then we find

$$S \approx 2 \times 10^{-27}.$$

Stimulated emission is thus very unlikely for equilibrium systems.

Another point worthy of note is that, for given values of N_2 (density of atoms in upper state E_2) and ρ_ν (density of photons) the rate of stimulated emission (B_{21}) is proportional to $1/\nu^3$. This follows from equation (A2.1.12*b*) since

$$B_{21} = \frac{A_{21}c^3}{8\pi h\nu^3}$$

and A_{21} is an atomic constant, representing the reciprocal of the spontaneous decay time.

This means that the higher the frequency the more difficult is laser action, for this depends upon stimulated emission. Ultraviolet, x-ray and γ-ray lasers present very special problems which, hopefully, will preclude the possibility of 'death-ray' weapons (x-rays and γ-rays are very damaging to living tissues).

However, we do wish to use lasers at lower frequencies, visible and infrared for example, for purposes of communication, display and measurement, and the equation for R_{21} tells us that the way to increase the stimulated emission is to increase the values of N_2 and ρ_ν.

We know that, in equilibrium, $N_2 < N_1$, from the form of the Boltzmann factor, and ρ_ν is given by equation (A2.1.8). Hence, we shall have to disturb the equilibrium to achieve significant levels of stimulated emission.

One way in which this can be done is to inject radiation at frequency ν, so that ρ_ν is increased above its equilibrium value. Suppose that this is done until the stimulated emission greatly exceeds the spontaneous emission (which does not, of course, depend upon ρ_ν), i.e. until:

$$N_2\rho_\nu B_{21} \gg N_2A_{21}.$$

The condition for this, clearly, is that

$$\rho_\nu \gg \frac{A_{21}}{B_{21}}$$

which, from equation (A2.1.12*b*), means that:

$$\rho_\nu \gg \frac{8\pi h\nu^3}{c^3}.$$

However, increasing ρ_ν does also increase the stimulated absorption. In fact, equation (A2.1.10) becomes, when ρ_ν is large:

$$N_1\rho_\nu B_{12} = N_2\rho_\nu B_{21}.$$

But we also know from equations (A2.1.12*a*) and (A2.1.12*b*) that $B_{12} = B_{21}$; hence $N_1 = N_2$ under these conditions. In other words, an incoming photon at frequency ν is just as likely to cause a downward transition (stimulated emission) as it is an upward one (stimulated absorption). Hence we cannot increase the population N_2 above that of N_1 simply by pumping more radiation, at frequency ν, into the system. Clearly, we must change tack if we are to enhance the stimulated emission and produce a laser.

Consider a three-level rather than a two-level system (figure A2.1.4(*a*)). Suppose that light at frequency ν_{13} is injected into this system, so that there is a large amount of stimulated absorption from E_1 to E_3. Spontaneous decays will occur from E_3 to E_2 and then $E_2 \rightarrow E_1$ with also $E_3 \rightarrow E_1$; but if the levels are chosen appropriately according to the quantum rules, the $E_3 \rightarrow E_2$ decay can be fast and the $E_2 \rightarrow E_1$ relatively much slower. Clearly, the result of this will be that atoms will accumulate in level E_2. Now the really important point is that, unlike the previous two-level case, atoms in level E_2 are immune from stimulated emission by photons at frequency ν_{13}. Hence we can now increase the number of atoms in level E_2, at the expense of those in E_1, by increasing the intensity of the radiation at frequency ν_{13}.

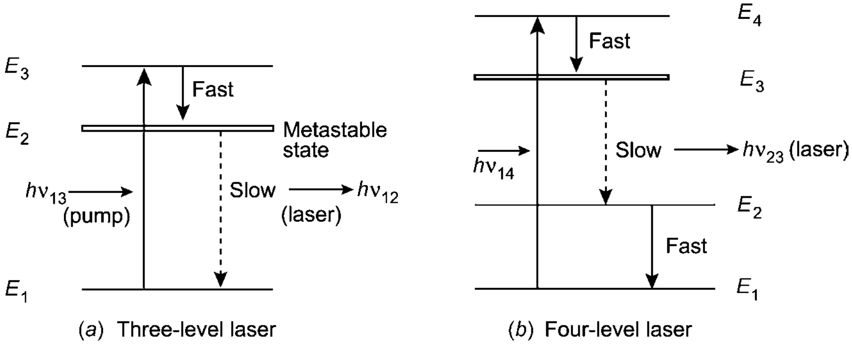


Figure A2.1.4. Energy-level diagrams for laser action.

We can thus soon ensure that:

$$N_2 > N_1$$

and we have an 'inverted population' (i.e. more atoms in a higher energy state than a lower one) as a result of the 'pump' at frequency ν_{13}. This inverted population can now be exploited to give optical amplification at frequency ν_{12}.

Let us quantify this amplification via the rate equations we have developed. Suppose that photons at frequency ν_{12} are injected into the medium in a certain direction. These will meet the inverted population in energy state E_2 and will stimulate the downward transition $E_2 \rightarrow E_1$, producing more photons at frequency ν_{12} in so doing (this is, of course, the origin of the amplification). We assume, quite confidently, that the medium is being sufficiently strongly pumped for the stimulated photons to be well in excess of any spontaneous emission from E_2 to E_1. Now suppose that, under these conditions, the number of photons per unit volume when the injected radiation enters the system is p_{12}. Then the rate at which p_{12} increases will depend upon the difference between upward and downward transition rates between levels 1 and 2, and hence we write:

$$\frac{\mathrm{d}p_{12}}{\mathrm{d}t} = N_2 \rho_{\nu_{12}} B_{21} - N_1 \rho_{\nu_{12}} \mathrm{B}_{12}.$$

Now $\rho_{\nu_{12}}$ is the energy density of photons, hence

$$\rho_{\nu_{12}} = p_{12} h\nu_{12}.$$

Also we know that $B_{12} = B_{21}$ (from equations (A2.1.12*a*) and (A2.1.13) and thus:

$$\frac{1}{h\nu_{12}} \frac{\mathrm{d}\rho_{\nu_{12}}}{\mathrm{d}t} = B_{12} \rho_{\nu_{12}} (N_2 - N_1). \tag{A2.1.13}$$

We shall now write $\rho_{\nu_{12}}$ as ρ_ν to avoid cluttered equations and, integrating equation (A2.1.13):

$$\rho_\nu = \rho_{\nu,0} \exp[(N_2 - N_1)t]$$

where $\rho_{\nu,0} = \rho_\nu$ at $t = 0$.

If the injected wave is travelling at velocity c in the medium we can transfer to a distance parameter via $s = ct$ and obtain:

$$\rho_\nu = \rho_{\nu,0} \exp\left[\frac{h\nu}{c} B_{12} (N_2 - N_1) s\right].$$

This is to be compared with the standard loss/gain relation for propagation in an interactive medium, i.e.

$$I = I_0 \exp(gx)$$

and it is clear that the gain coefficient g can be identified as:

$$g = \frac{h\nu}{c} B_{12}(N_2 - N_1) \tag{A2.1.14a}$$

which is the gain coefficient for the medium (fractional increase in intensity level per unit length) and will be positive (i.e. gain rather than loss) provided that $N_2 > N_1$, as will be the case for an inverted population. Hence, this medium is an optical amplifier. The injected radiation at frequency ν_{12} receives gain from the optical pump of amount:

$$G = \frac{I}{I_0} = \exp(gs)$$

so that it increases exponentially with distance into the medium. Clearly g in equation (A2.1.14*a*) is proportional to $(N_2 - N_1)$. In a three-level system such as we are considering the lower level of the amplifying transition is the ground state, which is initially heavily populated. It follows that more than half the atoms must be excited by the pump before population inversion can be achieved $(N_2 > N_1)$. It is quite hard work for the pump to excite all these atoms. Consider, however, the *four*-level system shown in figure A2.1.4(*b*). Here the pump is at ν_{14}, there is a quick decay to level 3 and a slow one to levels 2 and 1. The decay from 2 to 1 is again fast. Clearly, the consequence of this is that it is relatively easy to provide level 3 with an inverted population over level 2, since level 2 was not well populated in the first place (being above the ground state), and atoms do not accumulate there since it decays quickly to ground. Hence we can ensure that:

$$N_3 \gg N_2$$

with much less pump power than for $N_2 > N_1$ in the three-level case. The amplification at ν_{32} is thus much more efficient, and the four-level system makes for a more efficient amplifier.

The laser structure

Having arranged for efficient amplification to take place in a medium it is a relatively straightforward matter to turn it into an oscillator, i.e. a laser source. To do this for any amplifier it is necessary to provide positive feedback, i.e. to feed some of the amplified output back into the amplifier in reinforcing phase. This is done by placing parallel mirrors at each end of a box containing the medium, to form a Fabry–Perot cavity. The essential physics of this process is that any given photon at ν_{12} will be bounced back and forth between the mirrors, stimulating the emission of other such photons as it does so, whereas without the mirrors it would make only one such pass.

An important condition for any system to oscillate under these circumstances is that the gain should be in excess of the loss for each cycle of oscillation. The total loss for a photon executing a double passage of the cavity (figure A2.1.5) will depend not only on the loss per unit length in the medium (due to scattering, excitations to other states, wall losses, etc) but also on the losses at the mirrors and, it must be remembered that one of the mirrors has to be a partial mirror in order to let some of the light out otherwise we could not use the laser oscillator as a source! Hence the condition for oscillation is:

$$\frac{I_f}{I_i} = R_1 R_2 \exp[(g - \alpha)2l] > 1 \tag{A2.1.14b}$$

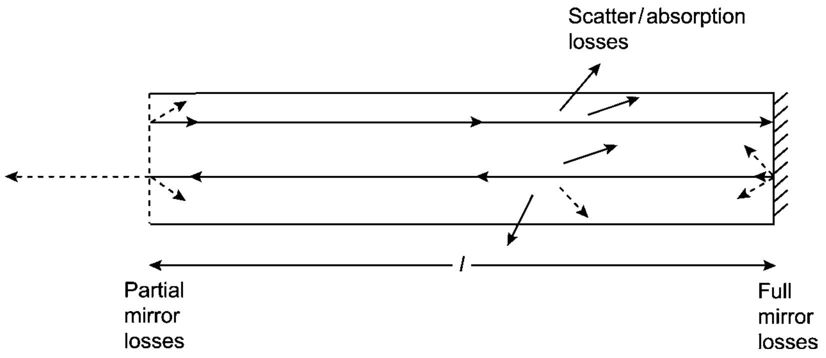


Figure A2.1.5. Loss mechanisms in laser cavity.

where I_f and I_i refer to the final and initial intensities for the double passage of the cavity, R_1 and R_2 are the reflectivities for the two mirrors, respectively, α is the loss per unit length in the medium and l is the cavity length. (The factor 2 in the exponential refers to the double passage of the photon.)

One further word of warning: the value of g must correspond to the population inversion whilst oscillation is taking place, not the value before the feedback is applied. Clearly the value of N_2/N_1 will be very different, once stimulated emission starts to occur, from its value when the system is simply being pumped into its inverted state. This has implications for pumping rates and the balancing of rate equations which we shall not pursue: the principles, hopefully, are clear, however.

The simple arrangement of a pumped medium lying between two parallel mirrors (one partial) will, under the correct pump conditions, therefore lead to radiation emerging with the following properties:

(a) Narrow linewidth, since only one energy of transition is involved in the laser action; and the mirrors, if wavelength selective, will block any spontaneous light which is emitted in addition.

(b) The output direction of the light will be exactly normal to the (accurately parallel) planes of the mirrors and thus will be highly collimated in one direction.

(c) When a photon is emitted via stimulated emission by another photon, it is emitted with the same phase as the original photon (remember the driving force/resonating system analogy), thus all the laser photons are locked in phase: we have coherent light (within the limitations only of the linewidth of the transition).

(d) The light can be very intense since all the 'light amplification by stimulated emission of radiation' from a long length of medium with small cross-sectional area can be collimated into one direction.

The above important features summarize the basic properties of laser light: it is pure (in wavelength and phase), intense, well-collimated light. It is thus easy to control and modulate; it is a powerful tool.

In order to enhance its usefulness as a tool there are two quite simple additions which can be made to the basic design: The Fabry–Perot cavity formed by the two parallel mirrors will possess defined longitudinal 'modes'. Waves propagating in opposite directions within the cavity, normal to the mirrors, will interfere and reinforce to give rise to an allowable stable mode only when

$$2L = m\lambda$$

where L is the length of the cavity and m is an integer.

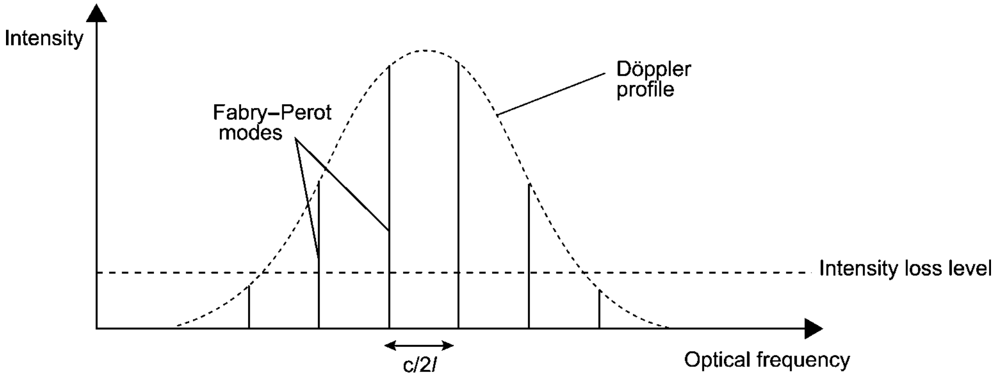


Figure A2.1.6. Laser-cavity spectrum.

From this we can also write

$$\lambda = \frac{2L}{m}; \quad f = \frac{cm}{2L}.$$

At all other wavelengths there is destructive interference. Now the stimulated emission occurs over a small range of wavelengths. This range is determined by the spectral width of the downward transition. The width depends upon a number of factors but primarily (unless cooled to very low temperatures) on the Doppler shift caused by the thermal motion of the molecules. Clearly, at any given time, some molecules will be moving towards the stimulating photon and others away, leading to a spread of Doppler shifts around the central line for the stationary molecule (at absolute zero of temperature!).

The output spectrum of the laser light is thus the result of combining these two features, as shown in figure A2.1.6. Here we can see the Fabry–Perot mode structure enveloped by the natural linewidth of the transition. In order to fix ideas somewhat, let us insert some real numbers into this. Suppose we have a He–Ne gas laser with length 0.5 m. Since, in a gas at less than atmospheric pressure, we have $c \sim 3 \times 10^8 \, \mathrm{ms}^{-1}$, we see that:

$$\frac{c}{2L} = 300 \, \mathrm{MHz}$$

which is the separation of the modes along the frequency axis. Now the Doppler line width of the 632.8 nm transition at 300 K is $\sim$ 1.5 GHz; hence the number of modes within this width is

$$\frac{1.5 \times 10^9}{3 \times 10^8} \sim 5$$

so that we have just five modes in the output spectrum.

So far we have dealt only with longitudinal modes; but off-axis rays also may interfere (figure A2.1.7). The reinforcement condition now depends also on the angle which the ray makes with the long axis, and the result is a variation in intensity over the cross-section of the cavity, and thus over the cross-section of the output laser beam (figure A2.1.7). (The notation used to classify these variations will be described in more detail when we deal with wave guiding (section A2.1.5) but TEM stands for 'transverse electromagnetic' and the two suffixes refer to the number of minima in the pattern in the horizontal and vertical directions, respectively.)

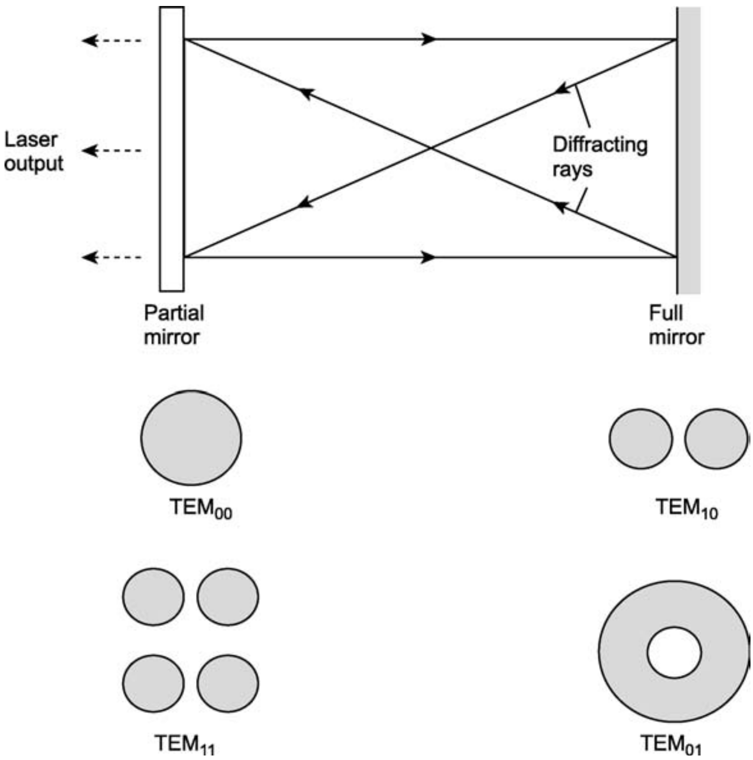


Figure A2.1.7. Transverse cavity modes.

Mode-locking

Let us return now to the longitudinal mode structure of the laser cavity. Normally, these longitudinal modes are entirely independent, since they result from wholly independently-acting interference conditions. Suppose, however, that they were to be locked into a constant phase relationship. In that case we would have a definite relationship, in the frequency domain, between the phases and the amplitudes of the various components of the frequency spectrum. If we were to translate those relationships into the time domain, by means of a Fourier transform, the result would be a series of pulses spaced by the reciprocal of the mode frequency interval, with each pulse shape the Fourier transform of the mode envelope (figure A2.1.8). All we are really saying here, in physical terms, is that if each frequency component bears a constant phase relationship to all the others, when all frequency components are superimposed, then there will be certain points in time where maxima occur (the peaks of the pulses) and others where minima occur (the troughs between pulses). If there is no fixed phase relationship between components both maxima and minima are ‘washed out’ into a uniform-level, randomized continuum.

Now a series of evenly spaced pulses is often a very useful form of laser output, so how can it be achieved?

We must lock the phases of the longitudinal modes. One way of doing this is to include, within the cavity, an amplitude modulator, and then modulate (not necessarily sinusoidally) the amplitudes of the modes at just the mode frequency interval, $c/2L$. Then, each mode generates a

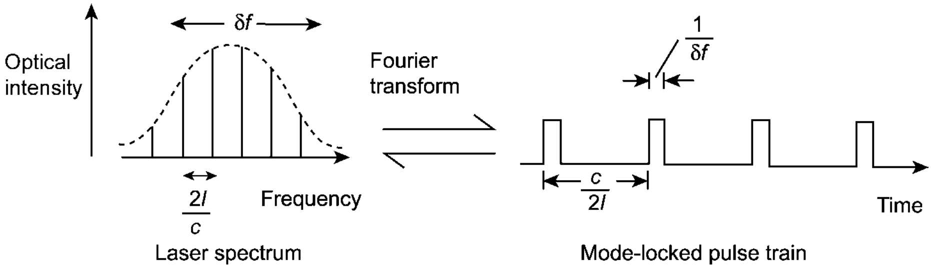


Figure A2.1.8. Mode-locking Fourier transform: spectrum/pulse train.

series of sidebands at frequencies $mc/2L$, which corresponds to the frequencies of the other modes. The result of this is that all the modes are 'pulled' mutually into phase by the driving forces at the other frequencies, and complete phase locking occurs. The inserted modulator thus has the effect of producing, from the laser output, a pulse stream with pulse repetition rate $2L/c$. For example, with the He–Ne laser quoted in section the laser structure the repetition rate is 300 MHz, and each pulse has a width

$$\sim \frac{1}{1.5\times 10^9} \sim 0.67\,\text{ns} \quad \text{(see figure A2.1.8).}$$

The laser is now said to be 'mode-locked' and the pulse stream is a set of 'mode-locked pulses'. Sometimes when a laser is being pumped quite hard and the output levels are high, the laser will 'self-mode lock'. This is due to the fact that the medium has been driven into the nonlinear regime (see section A2.1.7) and the modes generate their own harmonics as a result of the induced optical nonlinearities. Clearly, this will depend upon the medium as well as the driving level, since it will depend on which particular nonlinear threshold is exceeded by the pumping action.

Q-switching

The 'Q' or 'quality factor' of an oscillator refers to its purity, or 'sharpness of resonance'. The lower the loss in an oscillator the narrower is its resonance peak and the longer it will oscillate on its own after a single driving impulse. The equivalent quantity in a Fabry–Perot cavity (an optical oscillator) is the 'finesse', and the two quantities are directly related. From these ideas we can readily understand that if the loss in a resonator is varied then so is its 'Q'.

Suppose we have a laser medium sitting in its usual Fabry–Perot cavity but with a high loss; this means that a large fraction of the light power oscillating between the mirrors is lost per pass: we might, for example, have one of the mirrors with very low reflectivity.

Now the oscillator can only oscillate if the gain which the light receives per double pass between the mirrors exceeds the loss per double pass (section A2.1.2.4.2), and we shall suppose that the loss is very high, so that as we pump more and more molecules of the medium up into the excited state of the inverted population, the loss still exceeds the gain for as hard as our pump source can work. The result is that the inversion of the population becomes very large indeed, for there are very few photons to cause stimulated emission down to the lower state—they are all being lost by other means (e.g. a poor mirror at one end). Having achieved this very highly inverted population suppose that the loss is now suddenly reduced by means of an intercavity switch ('Q' switch) by, for example, speedily rotating to a high-reflectivity mirror (figure A2.1.9). The result is that there is suddenly an enormous number of photons to depopulate the inverted population,

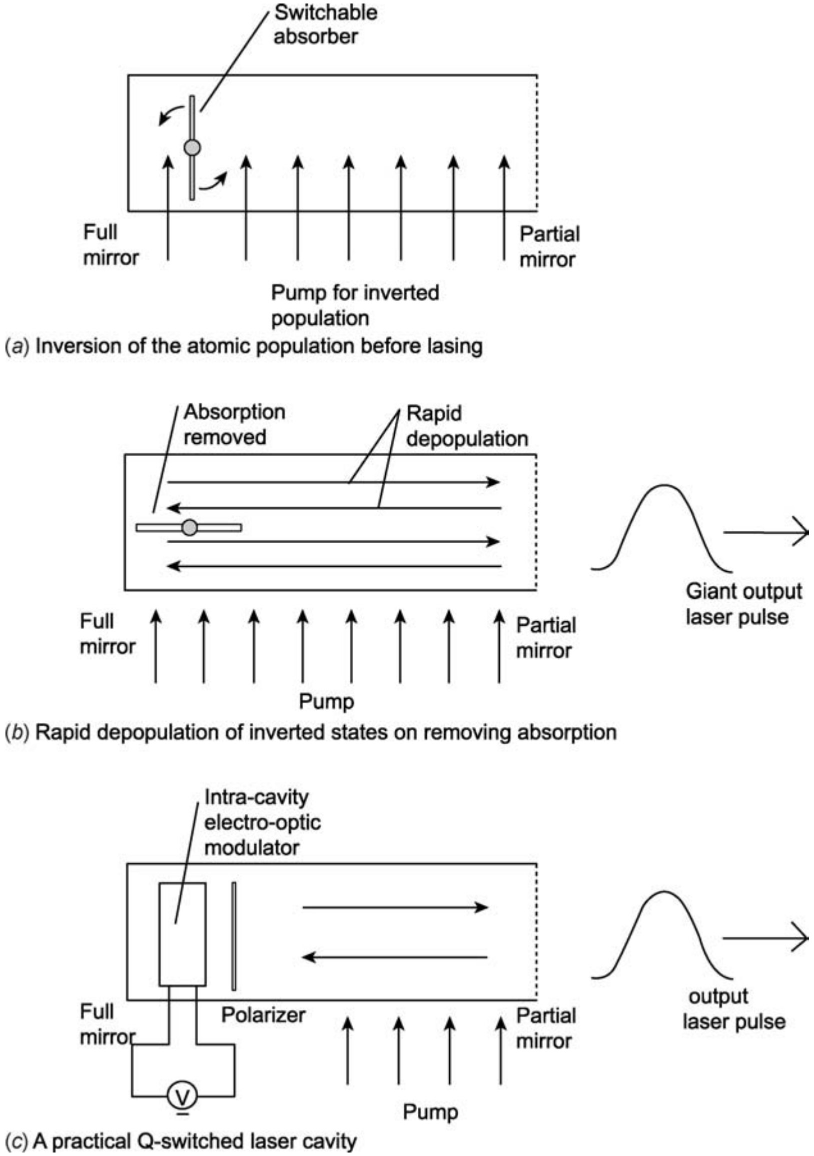

Switchable
absorber
Full
mirror
Partial
mirror
Pump for inverted
population
(a) Inversion of the atomic population before lasing
Absorption
removed
Rapid
depopulation
Giant output
laser pulse
Full
mirror
Partial
mirror
Pump
(b) Rapid depopulation of inverted states on removing absorption
Intra-cavity
electro-optic
modulator
output
laser pulse
Full
mirror
Polarizer
Partial
mirror
V
Pump
(c) A practical Q-switched laser cavity

Figure A2.1.9. Q-switching.

which then rapidly de-excites to emit all its accumulated energy in one giant laser pulse—the Q-switched pulse. Thus we have the means by which very large energy, very high intensity pulses can be obtained, albeit relatively infrequently (~25 pps).

Three further points should be noted concerning Q-switching:

(a) At the end of the pulse the lasing action ceases completely, since the large number of photons suddenly available completely depopulates the upper laser state.

(b) The switching to the low loss condition must take place in a time which is small compared with the stimulated depopulation time of the upper state, so as to allow the pulse to build up very quickly.

(c) The pumping rate must be large compared with the spontaneous decay rate of the upper state so as to allow a large population inversion to occur.

Q-switching can produce pulses with several millijoules of energy with only a few nanoseconds duration. Thus, peak powers of several megawatts can result. Such powers take most media into their nonlinear regimes (many will be evaporated!), so Q-switching is very useful for studying the nonlinear optical effects which will be considered in section A2.1.7.

Both mode-locking and Q-switching require intracavity modulation devices. These can take a variety of forms.

A2.1.2.5 Summary

In this section, we have examined those aspects of the physics of radiation relevant to optoelectronics. It is necessary to understand the essential quantum nature of radiation in order to appreciate fully the way in which light interacts with matter and, in particular, the action of the laser in providing a source of pure light. It was the invention of the laser (in 1960) which launched the subject of optoelectronics.

A2.1.3 Optical polarization

A2.1.3.1 Introduction

The essential idea of optical polarization was introduced in chapter A1.1 but we must now consider this important topic in more detail. We know that the electric and magnetic fields, for a freely propagating light wave, lie transversely to the propagation direction and orthogonally to each other.

Normally, when discussing polarization phenomena, we fix our attention on the electric field, since it is this which has the most direct effect when the wave interacts with matter.

In saying that an optical wave is polarized we are implying that the direction of the optical field is either constant or is changing in an ordered, prescribable manner. In general, the tip of the electric vector circumscribes an ellipse, performing a complete circuit in a time equal to the period of the wave, or in a distance of one wavelength. Clearly, the two parameters are equivalent in this respect.

As is well known, linearly polarized light can conveniently be produced by passing any light beam through a sheet of polarizing film. This is a material which absorbs light of one linear polarization (the ‘acceptance’ direction) to a much smaller extent (~1000 times) than the orthogonal polarization, thus, effectively, allowing just one linear polarization state to pass. The material’s properties result from the fact that it consists of long-chain polymeric molecules aligned in one direction (the acceptance direction) by stretching a plastic, and then stabilizing it. Electrons can move more easily along the chains than transversely to them, and thus the optical wave transmits easily only when its electric field lies along this acceptance direction. The material is cheap and allows the use of large optical apertures. It thus provides a convenient means whereby, for example, a specific linear polarization state can be defined; this state then provides a ready polarization reference which can be used as a starting point for other manipulations. In order to study these manipulations and other aspects of polarization optics, we shall begin by looking more closely at the polarization ellipse.

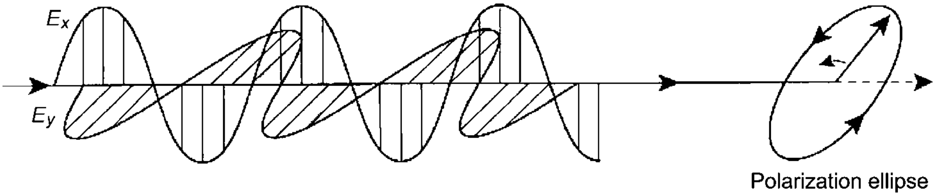


Figure A2.1.10. Components for an elliptically-polarized wave.

A2.1.3.2 The polarization ellipse

In chapter A1.1, the most general form of polarized light wave propagating in the *Oz* direction was derived from the two linearly polarized components in the *Ox* and *Oy* directions (figure A2.1.10):

$$E_x = e_x \cos(\omega t - kz + \delta_x) \quad E_y = e_y \cos(\omega t - kz + \delta_y). \tag{A2.1.15a}$$

If we eliminate $(\omega t - kz)$ from these equations we obtain the expression:

$$\frac{E_x^2}{e_x^2} + \frac{E_y^2}{e_y^2} + \frac{2E_xE_y}{e_xe_y}\cos(\delta_y - \delta_x) = \sin^2(\delta_y - \delta_x) \tag{A2.1.15b}$$

which is the ellipse (in the variables E_x, E_y) circumscribed by the tip of the resultant electric vector at any one point in space over one period of the combined wave. This can only be true, however, if the phase difference $(\delta_y - \delta_x)$ is constant in time, or, at least, changes only slowly when compared with the speed of response of the detector. In other words, we say that the two waves must have a large mutual 'coherence'. If this was not so then relative phases and hence resultant field vectors would vary randomly within the detector response time, giving no ordered pattern to the behaviour of the resultant field and thus presenting to the detector what would be, essentially, unpolarized light.

Assuming that the mutual coherence is good, we may investigate further the properties of the polarization ellipse.

Note, firstly, that the ellipse always lies in the rectangle shown in figure A2.1.11, but that the axes of the ellipse are not parallel with the original x, y directions.

The ellipse is specified as follows: with e_x, e_y, δ $(= \delta_y - \delta_x)$ known, then we define $\tan\beta = e_y/e_x$.

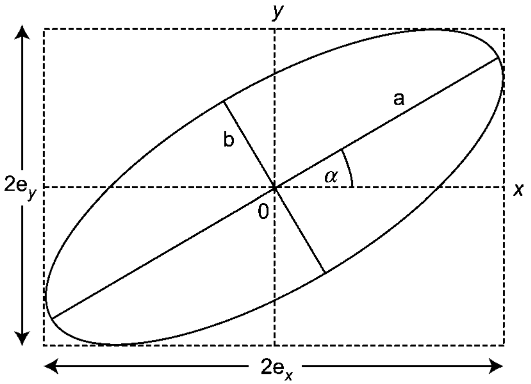


Figure A2.1.11. The polarization ellipse.

The orientation of the ellipse, α, is given by:

$$\tan 2\alpha = \tan 2\beta \cos \delta.$$

Semi-major and semi-minor axes a, b are given by:

$$e_x^2 + e_y^2 = a^2 + b^2 \sim I.$$

The ellipticity of the ellipse (e) is given by $e = \tan\chi = \pm\, b/a$ (the sign determines the sense of the rotation) where:

$$\sin 2\chi = -\sin 2\beta \sin \delta.$$

We should note also that the electric field components along the major and minor axes are always in quadrature (i.e. $\pi/2$ phase difference, the sign of the difference depending on the sense of the rotation).

Linear and circular states of polarization may be regarded as special cases where the polarization ellipse degenerates into a straight line or a circle, respectively.

A linear state is obtained with the components in equation (A2.1.15a) when either:

$$\left.\begin{aligned} e_x &= 0 \\ e_y &\neq 0 \end{aligned}\right\} \quad \text{linearly polarized in } Oy \text{ direction}$$

$$\left.\begin{aligned} e_x &\neq 0 \\ e_y &= 0 \end{aligned}\right\} \quad \text{linearly polarized in } Ox \text{ direction}$$

or,

$$\delta_y - \delta_x = m\pi$$

where m is an integer. In this latter case, the direction of polarization will be at an angle:

$$+\tan^{-1}\left(\frac{e_y}{e_x}\right) \qquad m \text{ even}$$

$$-\tan^{-1}\left(\frac{e_y}{e_x}\right) \qquad m \text{ odd}$$

with respect to the Ox axis.

A circular state is obtained when

$$e_x = e_y$$

and

$$(\delta_y - \delta_x) = (2m + 1)\pi/2$$

i.e. in this case the two waves have equal amplitudes and are in phase quadrature. The waves will be right-hand circularly polarized when m is even and left-hand circularly polarized when m is odd.

Light can become polarized as a result of the intrinsic directional properties of matter: either the matter which is the original source of the light, or the matter through which the light passes. These intrinsic material directional properties are the result of directionality in the bonding which holds together the atoms of which the material is made. This directionality leads to variations in the response of the material according to the direction of an imposed force, be it electric, magnetic or mechanical.

The best known manifestation of directionality in solid materials is the crystal, with the large variety of crystallographic forms, some symmetrical, some asymmetrical. The characteristic shapes which we associate with certain crystals result from the fact that they tend to break preferentially along certain planes known as cleavage planes, which are those planes between which atomic forces are weakest.

It is not surprising, then, to find that directionality in a crystalline material is also evident in the light which it produces, or is impressed upon the light which passes through it.

In order to understand the ways in which we may produce polarized light, control it and use it, we must make a gentle incursion into the subject of crystal optics.

A2.1.3.3 Crystal optics

Light propagates through a material by stimulating the elementary atomic dipoles to oscillate and thus to radiate. In our previous discussions the forced oscillation was assumed to take place in the direction of the driving electric field, but in the case of a medium whose physical properties vary with direction, an anisotropic medium, this is not necessarily the case. If an electron in an atom or molecule can move more easily in one direction than another, then an electric field at some arbitrary angle to the preferred direction will move the electron in a direction which is not parallel with the field direction (figure A2.1.12). As a result, the direction in which the oscillating dipole's radiation is maximized (i.e. normal to its oscillation direction) is not the same as that of the driving wave.

The consequences, for the optics of anisotropic media, of this simple piece of physics are complex.

Immediately we can see that the already-discussed (see chapter A1.1) relationship between the electric displacement $\boldsymbol{D}$ and the electric field $\boldsymbol{E}$, for an *isotropic* (i.e. no directionality) medium:

$$\boldsymbol{D} = \varepsilon_R \varepsilon_0 \boldsymbol{E}$$

must be more complex for an anisotropic medium; in fact the relation must now be written in the form (for any, arbitrary three orthogonal directions Ox, Oy, Oz):

$$D_x = \varepsilon_0(\varepsilon_{xx}E_x + \varepsilon_{xy}E_y + \varepsilon_{xz}E_z) \quad D_y = \varepsilon_0(\varepsilon_{yx}E_x + \varepsilon_{yy}E_y + \varepsilon_{yz}E_z) \quad D_z = \varepsilon_0(\varepsilon_{zx}E_x + \varepsilon_{zy}E_y + \varepsilon_{zz}E_z).$$

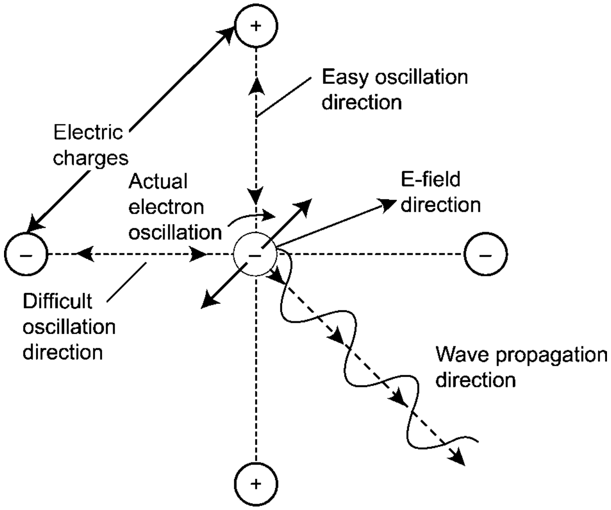

+
Easy oscillation direction
Electric charges
Actual electron oscillation
E-field direction
–
–
–
Difficult oscillation direction
Wave propagation direction
+

Figure A2.1.12. Electron response to an electric field in a dielectric medium.

Clearly what is depicted here is an array which describes the various electric field susceptibilities in the various directions within the crystal: ε_{ij} (a scalar quantity) is a measure of the effect which an electric field in direction j has in direction i within the crystal, i.e. the ease with which it can move electrons in that direction and thus create a dipole moment.

The array can be written in the abbreviated form

$$D_i = \varepsilon_0 \varepsilon_{ij} E_j \quad (i, j = x, y, z)$$

and ε_{ij} is now a *tensor* known, in this case, as the permittivity tensor. A tensor is a physical quantity which characterizes a particular physical property of an anisotropic medium, and takes the form of a matrix. Clearly $\boldsymbol{D}$ is not now (in general) parallel with $\boldsymbol{E}$, and the angle between the two also will depend upon the direction of $\boldsymbol{E}$ in the material.

Now it can be shown from energy considerations that the permittivity tensor is symmetrical, i.e. $\varepsilon_{ij} = \varepsilon_{ji}$. Also, symmetrical tensors can be cast into their diagonal form by referring them to a special set of axes (the principal axes) which are determined by the crystal structure [2]. When this is done, we have:

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \varepsilon_0 \begin{pmatrix} \varepsilon_{xx} & 0 & 0 \\ 0 & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}.$$

The new set of axes, Ox, Oy, Oz, is now this special set.

Suppose now that $\boldsymbol{E} = E_x \boldsymbol{i}$, i.e. we have, entering the crystal, an optical wave whose $\boldsymbol{E}$ field lies in one of these special crystal directions.

In this case, we simply have:

$$D_x = \varepsilon_0 \varepsilon_{xx} E_x$$

as our tensor relation and e_{xx} is, of course, a scalar quantity. In other words, we have $\boldsymbol{D}$ parallel with $\boldsymbol{E}$, just as for an isotropic material, and the light will propagate, with refractive index $e_{xx}^{1/2}$, perfectly normally. Furthermore, the same will be true for:

$$\boldsymbol{E} = E_y \boldsymbol{j}, \quad (\text{refractive index } e_{yy}^{1/2})$$

$$\boldsymbol{E} = E_z \boldsymbol{k}, \quad (\text{refractive index } e_{zz}^{1/2}).$$

Before going further we should note an important consequence of all this: the refractive index varies with the direction of $\boldsymbol{E}$. If we have a wave travelling in direction Oz, its velocity now will depend upon its polarization state: if the wave is linearly polarized in the Ox direction it will travel with velocity $c_0/e_{xx}^{1/2}$ while if it is linearly polarized in the Oy direction its velocity will be $c_0/e_{yy}^{1/2}$. Hence the medium is offering two refractive indices to the wave travelling in this direction: we have the phenomenon known as double refraction or 'birefringence'. A wave which is linearly polarized in a direction at 45° to Ox will split into two components, linearly polarized in directions Ox and Oy, the two components travelling at different velocities. Hence the phase difference between the two components will steadily increase and the composite polarization state of the wave will vary progressively from linear to circular and back to linear again.

This behaviour is a direct consequence of the basic physics which was discussed earlier: it is easier, in the anisotropic crystal, for the electric field to move the atomic electrons in one direction than in another. Hence, for the direction of easy movement, the light polarized in this direction can travel faster than when it is polarized in the direction for which the movement is more sluggish. Birefringence is a long word, but the physical principles which underlie it really are very simple. It follows from these discussions that an anisotropic medium may be characterized by means of three refractive indices, corresponding to polarization directions along Ox, Oy, Oz, and that these will have values $e_{xx}^{1/2}$, $e_{yy}^{1/2}$, $e_{zz}^{1/2}$, respectively. We can use this information to determine the refractive index (and thus the velocity) for a wave in any direction with any given linear polarization state.

To do this we construct an 'index ellipsoid' or 'indicatrix', as it is sometimes called (see figure A2.1.13), from the form of the permittivity tensor for any given crystal. This ellipsoid has the following important properties.

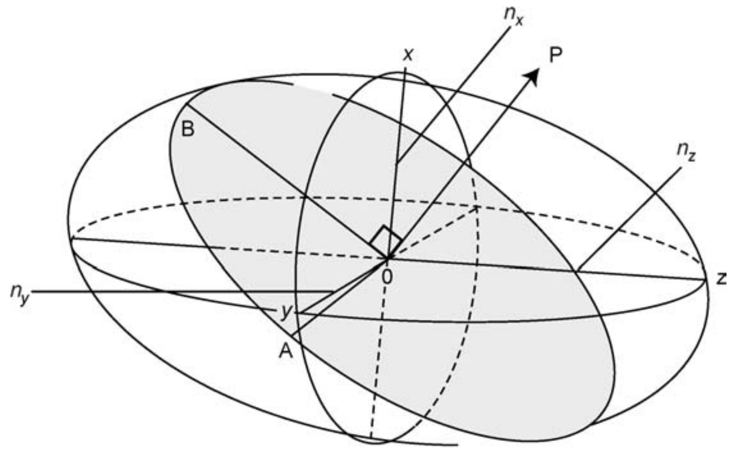
Suppose that we wish to investigate the propagation of light, at an arbitrary angle to the crystal axes (polarization as yet unspecified). We draw a line, OP, corresponding to this direction within the index ellipsoid, passing through its centre O (figure A2.1.13). Now we construct the plane, also passing through O, which lies at right angles to the line. This plane will cut the ellipsoid in an ellipse. This ellipse has the property that the directions of its major and minor axes define the directions of linear polarization for which $\boldsymbol{D}$ and $\boldsymbol{E}$ are parallel for this propagation direction, and the lengths of these axes OA and OB are equal to the refractive indices for these polarizations. Since these two linear polarization states are the only ones which propagate without change of polarization form for this crystal direction, they are sometimes referred to as the 'eigenstates' or 'polarization eigenmodes' for this direction, conforming to the matrix terminology of eigenvectors and eigenvalues.

The propagation direction we first considered, along Oz, corresponds, to one of the axes of the ellipsoid, and the two refractive indices $e_{xx}^{1/2}$ and $e_{yy}^{1/2}$ are the lengths of the other two axes in the central plane normal to Oz.

The refractive indices $e_{xx}^{1/2}$, $e_{yy}^{1/2}$, $e_{zz}^{1/2}$ are referred to as the *principal* refractive indices and we shall henceforth denote them n_x, n_y, n_z. $OxOy$, $OyOz$ and $OzOx$ are the principal planes.

Several other points are very well worth noting. Suppose, firstly, that

$$n_x > n_y > n_z.$$



OA and OB represent the linearly polarized eigenstates for the direction OP

Figure A2.1.13. The index ellipsoid.

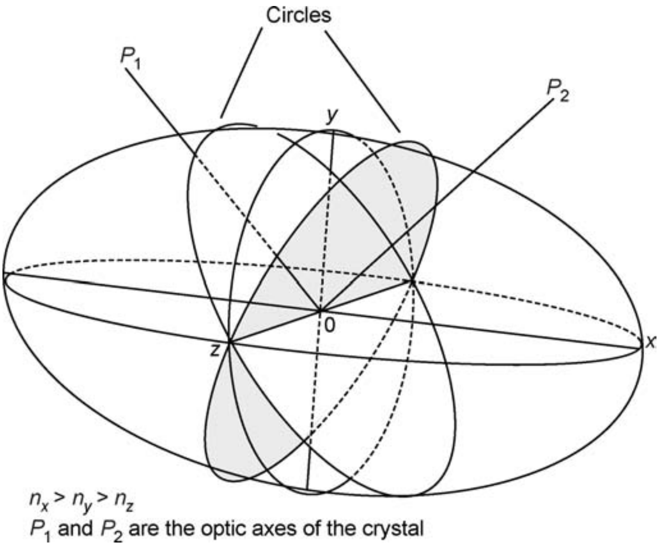


Figure A2.1.14. The ellipsoid for a biaxial crystal.

It follows that there will be a plane which contains *Oz* for which the two axes of interception with the ellipsoid are equal (figure A2.1.14). This plane will be at some angle to the *yz* plane and will thus intersect the ellipsoid in a circle. This means, of course, that, for the light propagation direction corresponding to the normal to this plane, all polarization directions have the same velocity; there is no double refraction for this direction. This direction is an *optic axis* of the crystal and there will, in general, be two such axes, since there must also be such a plane at an equal angle to the *yz* plane on the other side (see figure A2.1.14). Such a crystal with two optic axes is said to be biaxial.

Suppose now that:

$$n_x = n_y = n_\mathrm{o} \quad \text{(say), the `ordinary' index}$$

and

$$n_z = n_\mathrm{e} \quad \text{(say), the `extraordinary' index.}$$

In this case one of the principal planes is a circle and it is the only circular section (containing the origin) which exists. Hence, in this case there is only one optic axis, along the *Oz* direction. Such crystals are said to be uniaxial (figure A2.1.15). The crystal is said to be positive when $n_\mathrm{e} > n_\mathrm{o}$ and negative when $n_\mathrm{e} < n_\mathrm{o}$. For example, quartz is a positive uniaxial crystal, and calcite a negative uniaxial crystal. These features are, of course, determined by the crystal class to which these materials belong.

It is clear that the index ellipsoid is a very useful device for determining the polarization behaviour of anisotropic media.

A2.1.3.4 Circular birefringence

So far we have considered only linear birefringence, where two orthogonal linear polarization eigenstates propagate, each remaining linear, but with different velocities. Some crystals also exhibit circular birefringence. Quartz (again) is one such crystal and its circular birefringence derives from the

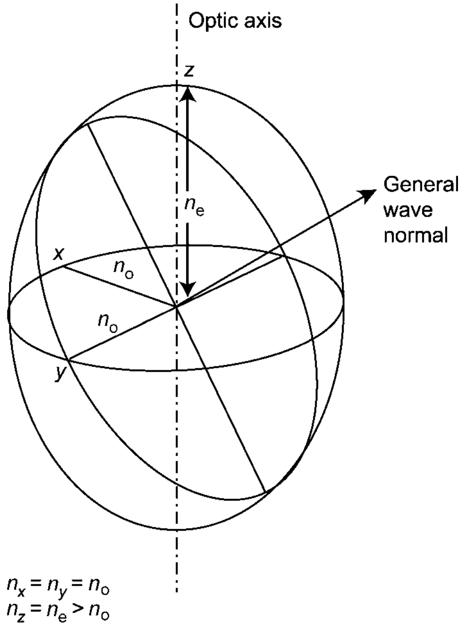


Figure A2.1.15. The ellipsoid for a positive uniaxial crystal.

fact that the crystal structure spirals around the optic axis in a right-handed (dextro-rotatory) or left-handed (laevo-rotatory) sense depending on the crystal specimen: both forms exist in nature.

It is not surprising to find, in view of this knowledge, and our understanding of the easy motions of electrons, that light which is right-hand circularly polarized (clockwise rotation of the tip of the electric vector as viewed by a receiver of the light) will travel faster down the axis of a matching right-hand spiralled crystal structure than left-hand circularly polarized light. We now have circular birefringence: the two circular polarization components propagate without change of form (i.e. they remain circularly polarized) but at different velocities. They are the circular polarization eigenstates for this case.

The term 'optical activity' has been traditionally applied to this phenomenon, and it is usually described in terms of the rotation of the polarization direction of a linearly polarized wave as it passes down the optic axis of an 'optically active' crystal. This fact is exactly equivalent to the interpretation in terms of circular birefringence, since a linear polarization state can be resolved into two oppositely rotating circular components (figure A2.1.16). If these travel at different velocities, a phase difference is inserted between them. As a result of this, when recombined, they again form a resultant which is linearly polarized but rotated with respect to the original direction (figure A2.1.16). Hence 'optical activity' is equivalent to circular birefringence. In general, both linear and circular birefringence might be present simultaneously in a material (such as quartz). In this case the polarization eigenstates which propagate without change of form (and at different velocities) will be elliptical states, the ellipticity and orientation depending upon the ratio of the magnitudes of the linear and circular birefringences, and on the direction of the linear birefringence eigen-axes within the crystal.

It should, again, be emphasised that only the polarization eigenstates propagate without change of form. All other polarization states will be changed into different polarization states by the action of the polarization element (e.g. a crystal component). These changes of polarization state are very useful in opto-electronics. They allow us to control, analyse, modulate and demodulate polarization information impressed upon a light beam, and to measure important directional properties relating to the medium

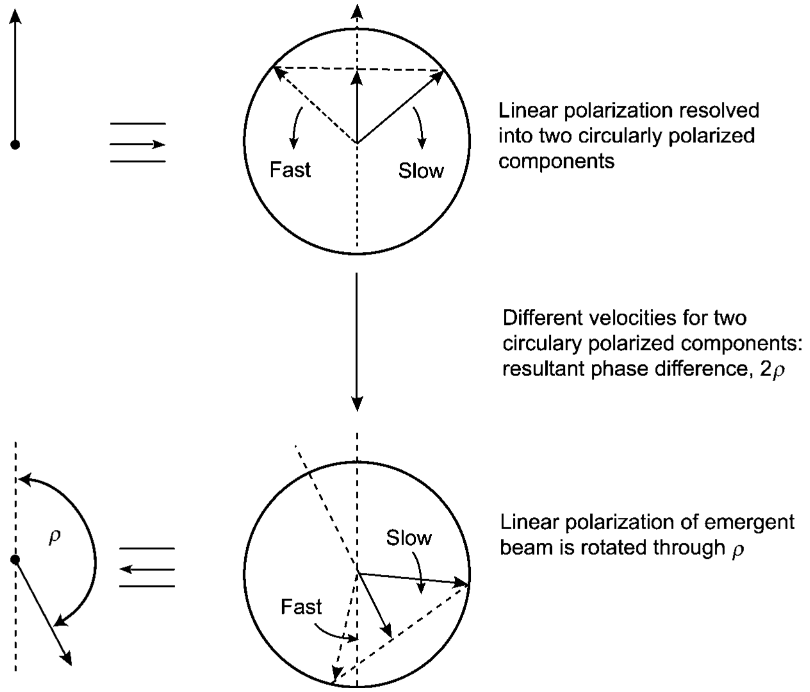

Fast
Slow
Linear polarization resolved into two circularly polarized components
Different velocities for two circulary polarized components: resultant phase difference, 2ρ
ρ
Slow
Fast
Linear polarization of emergent beam is rotated through ρ

Figure A2.1.16. Resolution of linear polarization into circularly-polarized components in circular birefringence (2ρ).

through which the light has passed. We must now develop a rigorous formalism to handle these more general polarization processes.

A2.1.3.5 Polarization analysis

As has been stated, with both linear and circular birefringence present, the polarization eigenstates (i.e. the states which propagate without change of form) for a given optical element are elliptical states, and the element is said to exhibit elliptical birefringence, since these eigenstates propagate with different velocities.

In general, if we have, as an input to a polarization-optical element, light of one elliptical polarization state, it will be converted, on emergence, into a different elliptical polarization state (the only exceptions being, of course, when the input state is itself an eigenstate). We know that any elliptical polarization state can always be expressed in terms of two orthogonal electric field components defined with respect to chosen axes Ox, Oy, i.e.

$$E_x = e_x \cos(\omega t - kz + \delta_x)$$
$$E_x = e_y \cos(\omega t - kz + \delta_y)$$

or, in complex exponential notation:

$$E_x = |E_x| \exp(\mathrm{i}\varphi_x); \quad \varphi_x = \omega t - kz + \delta_x$$
$$E_y = |E_y| \exp(\mathrm{i}\varphi_y); \quad \varphi_y = \omega t - kz + \delta_y.$$

When this ellipse is converted into another by the action of a lossless polarization element, the new ellipse will be formed from components which are linear combinations of the old, since it results from directional resolutions and rotations of the original fields. Thus these new components can be written:

$$E'_x = m_1 E_x + m_4 E_y$$
$$E'_y = m_3 E_y + m_2 E_y$$

or, in matrix notation

$$\mathbf{E}' = \mathbf{ME}$$

where

$$\mathbf{M} = \begin{pmatrix} m_1 & m_4 \\ m_3 & m_2 \end{pmatrix} \tag{A2.1.16}$$

and the m_n are, in general, complex numbers. **M** is known as a 'Jones' matrix after the mathematician who developed an extremely useful 'Jones calculus' for manipulations in polarization optics [3]. Now in order to make measurements of the input and output states in practice, we need a quick and convenient experimental method.

A convenient method for this practical determination is to use a linear polarizer and a quarter-wave plate, and to measure the light intensities for a series of fixed orientations of these elements.

Suppose that $I(\vartheta, \varepsilon)$ denotes the intensity of the incident light passed by the linear polarizer set at an angle ϑ to Ox, after the Oy component has been retarded by an angle ε as a result of the insertion of the quarter-wave plate with its axes parallel with O_x, O_y. We measure what are called the four Stokes parameters, as follows:

$$\begin{aligned}
S_0 &= I(0^\circ, 0) + I(90^\circ, 0) = e_x^2 + e_y^2 \\
S_1 &= I(0^\circ, 0) - I(90^\circ, 0) = e_x^2 - e_y^2 \\
S_2 &= I(45^\circ, 0) - I(135^\circ, 0) = 2e_x e_y \cos\delta \\
S_3 &= I\left(45^\circ, \frac{\pi}{2}\right) - I\left(135^\circ, \frac{\pi}{2}\right) = 2e_x e_y \sin\delta \\
\delta &= \delta_y - \delta_x.
\end{aligned}$$

If the light is 100% polarized, only three of these parameters are independent, since:

$$S_0^2 = S_1^2 + S_2^2 + S_3^2.$$

S_0 being the total light intensity.

If the light is only partially polarized, the fraction

$$\eta = \frac{S_1^2 + S_2^2 + S_3^2}{S_0^2}$$

defines the degree of polarization. In what follows we shall assume that the light is fully polarized ($\eta = 1$). It is easy to show [4a] that measurement of the S_n provides the ellipticity, e, and the orientation, α, of the polarization ellipse according to the relations:

$$\begin{aligned}
e &= \tan\chi \\
\sin 2\chi &= \frac{S_3}{S_0} \\
\tan 2\alpha &= \frac{S_2}{S_1}.
\end{aligned}$$

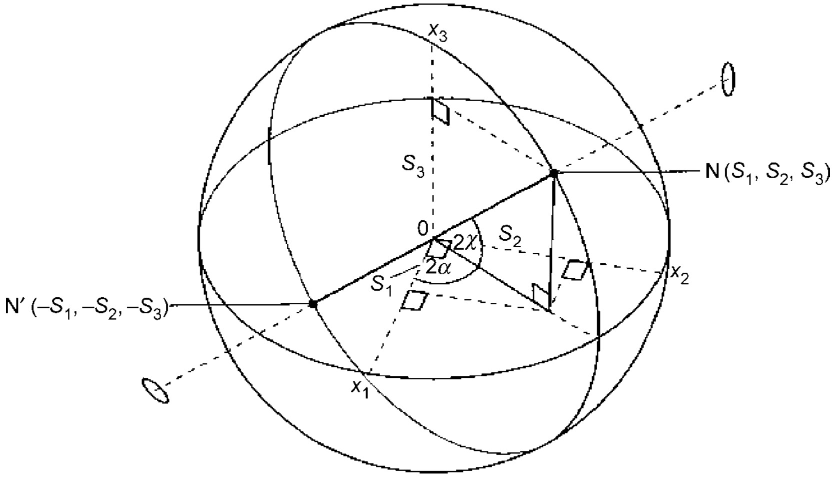


Figure A2.1.17. The Poincaré sphere: the eigenmode diameter (NN′).

Now, the above relations suggest a geometrical construction which provides a powerful and elegant means for description and analysis of polarization-optical phenomena. The Stokes parameters S_1, S_2, S_3 may be regarded as the Cartesian co-ordinates of a point referred to axes Ox_1, Ox_2, Ox_3. Thus every elliptical polarization state corresponds to a unique point in three-dimensional space. For a constant S_0 (lossless medium) it follows that all such points lie on a sphere of radius S_0: the Poincaré sphere (figure A2.1.17). The properties of the sphere are quite well known [5]. We can see that the equator will comprise the continuum of linearly polarized states, while the two poles will correspond to the two oppositely-handed states of circular polarization.

It is clear that any change, resulting from the passage of light through a lossless element, from one polarization state to another, corresponds to a rotation of the sphere about a diameter. Now any such rotation of the sphere may be expressed as a unitary 2 × 2 matrix **M**. Thus, the conversion from one polarization state **E** to another **E**′ may also be expressed in the form:

$$\mathbf{E}' = \mathbf{M}\mathbf{E}$$

or

$$\begin{pmatrix} E'_x \\ E'_y \end{pmatrix} = \begin{pmatrix} m_1 & m_4 \\ m_3 & m_2 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix}$$

i.e.

$$E'_x = m_1 E_x + m_4 E_y$$
$$E'_y = m_3 E_x + m_2 E_y$$

where

$$\mathbf{M} = \begin{pmatrix} m_1 & m_4 \\ m_3 & m_2 \end{pmatrix}$$

and **M** may be immediately identified with our previous **M** (equation (A2.1.16)).

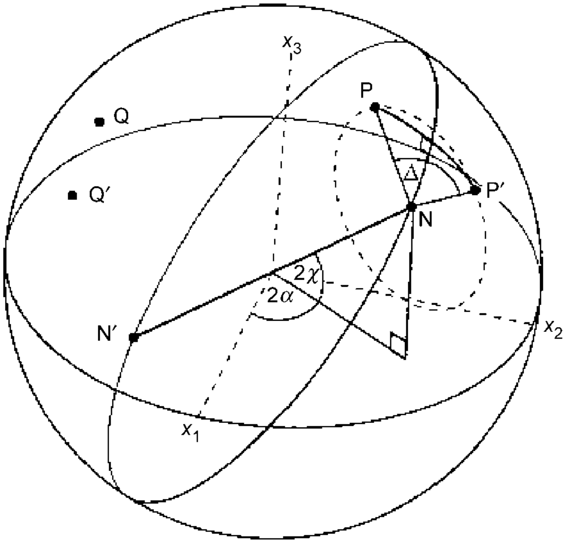


Figure A2.1.18. Rotation of the Poincaré sphere about an eigenmode diameter, NN′.

M is a Jones matrix [3] which completely characterizes the polarization action of the element and is also equivalent to a rotation of the Poincaré sphere.

The two eigenvectors of the matrix correspond to the eigenmodes (or eigenstates) of the element (i.e. those polarization states which can propagate through the element without change of form). These two polarization eigenstates lie at opposite ends of a diameter (NN′) of the Poincaré sphere, and the polarization effect of the element is to rotate the sphere about this diameter (figure A2.1.18) through an angle Δ which is equal to the phase which the polarization element inserts between its eigenstates.

The polarization action of the element may thus be regarded as that of resolving the input polarization state into the two eigenstates with appropriate amplitudes, and then inserting a phase difference between them before recombining to obtain the emergent state. Thus, a pure rotator (e.g. optically active crystal) is equivalent to a rotation about the polar axis, with the two oppositely handed circular polarizations as eigenstates. The phase velocity difference between these two eigenstates is a measure of the circular birefringence. Analogously, a pure linear retarder (such as a wave plate) inserts a phase difference between orthogonal linear polarizations which measures the linear birefringence. The linear retarder's eigenstates lie at opposite ends of an equatorial diameter.

It is useful for many purposes to resolve the polarization action of any given element into its linear and circular birefringence components. The Poincaré sphere makes it clear that this may always be done since any rotation of the sphere can always be resolved into two sub-rotations, one about the polar diameter and the other about an equatorial diameter.

From this brief discussion we can begin to understand the importance of the Poincaré sphere. It is a construction which converts all polarization actions into visualizable relationships in three-dimensional space.

To illustrate this point graphically let us consider a particular problem. Suppose that we ask what is the smallest number of measurements necessary to define completely the polarization properties of a given lossless polarization element, about which we know nothing in advance. Clearly, we must provide known polarization input states and measure their corresponding output states, but how many input/output pairs are necessary: one, two, more?

The Poincaré sphere answers this question easily. The element in question will possess two polarization eigenmodes and these will be at opposite ends of a diameter. We need to identify this diameter. We know that the action of the element is equivalent to a rotation of the sphere about this diameter, and through an angle equal to the phase difference which the element inserts between its eigenmodes. Hence, if we know one input/output pair of polarization states (NN′), we know that the rotation from the input to the output state must have taken place about a diameter which lies in the plane which perpendicularly bisects the line joining the two states (see figure A2.1.18). Two other input/output states (QQ′) will similarly define another such plane, and thus the required diameter is clearly seen as the common line of intersection of these planes.

Further, the phase difference Δ inserted between the eigenstates (i.e. the sphere's rotation angle) is easily calculated from either pair of states, once the diameter is known.

Hence the answer is that *two* pairs of input/output states will define completely the polarization properties of the element. Simple geometry has provided the answer. A good general approach is to use the Poincaré sphere to determine (visualize?) the nature of the solution to a problem, and then to revert to the Jones matrices to perform the precise calculations. Alternatively, some simple results in spherical trigonometry will usually suffice.

A2.1.3.6 Summary

In this section, we have looked closely at the directionality possessed by the optical transverse electric field, i.e. we have looked at optical polarization. We have seen how to describe it, to characterize it, to control it, to analyse it, and how, in some ways, to use it.

We have also looked at the ways in which the transverse electric and magnetic fields interact with directionalities (anisotropies) in material media through which the light propagates. In particular, we first looked at ways in which the interactions allow us to probe the nature and extent of the material directionalities, and thus to understand better the materials themselves.

Secondly, we looked briefly at the ways in which these material interactions allow us to control light: to modulate it, and perhaps to analyse it.

We shall find later that the knowledge we have gained bears upon more advanced phenomena, such as those which allow light to switch light and to process light, opening up a new range of possibilities in the world of very fast (femtosecond: 10^{-15} s) phenomena.

A2.1.4 Optical coherence

A2.1.4.1 Introduction

In dealing with interference and diffraction the assumption usually is made that each of the interfering waves bears a constant phase relationship to the others in both time and space. Such an assumption cannot be valid for all time and space intervals since the atomic emission processes which give rise to light are largely uncorrelated, except for the special case of laser emission. In this section, the topic of 'coherence' will be dealt with. Clearly, it will have a bearing on interference phenomena.

The coherence of a wave describes the extent to which it can be represented by a pure sine wave. A pure sine wave has infinite extension in space and time, and hence cannot exist in reality. Perfect coherence is thus unachievable, but it is nevertheless a valuable concept.

Coherence, in general, is a valuable concept because it is a measure of the constancy of the relationships between one part of a wave (in time and/or space) and another; and between one wave and another. This is why it is so important from the point of view of interference: a wave can only interfere with itself or with another wave (of the same polarization) to produce a sensible interference pattern if

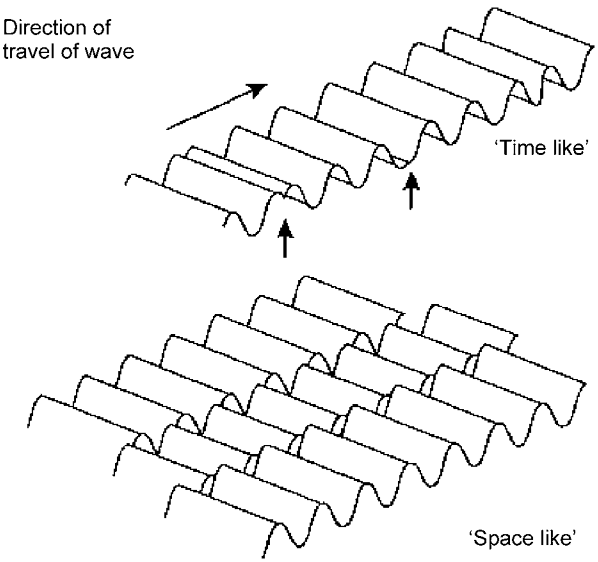


Figure A2.1.19. Illustrations for partial coherence.

the phases and amplitudes remain in constant relationship whilst the pattern is being sensed. Additionally, it is clear that if we wish to impose information on an optical wave by modulating one of its defining parameters (i.e. amplitude, phase, polarization or frequency) then the extent to which that information remains intact is mirrored by the extent to which the modulated parameter remains intact on the wave itself; thus coherence is an important parameter in respect of the optical wave's information-carrying capacity, and our ability generally to control and manipulate it.

Waves which quite accurately can be represented by sine waves for a limited period of time or in a limited region of space are called partially coherent waves. Figure A2.1.19 shows examples of time-like and space-like partial coherence.

A normal light source emits quanta of energy at random. Each quantum conveniently can be regarded as a finite wave train having angular frequency ω_0, say, and duration $2\tau_c$ (figure A2.1.20).

Fourier theory tells us that this wave train can be described in the frequency domain as a set of waves lying in the frequency range $\omega_0/2\pi \pm 1/\tau_c$. For a large number of randomly-emitted wave packets all the components at any given frequency will possess random relative phases. Spatial and temporal (time-like) coherence will thus only exist for, respectively, a distance of the order of the length of one

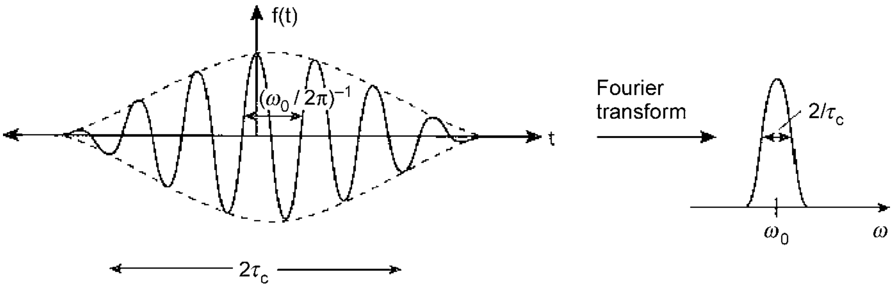


Figure A2.1.20. The optical wave packet.

packet ($2c\tau_c$) and a time of the order of its duration ($2\tau_c$). If the wave packets were of infinitely short duration (δ-function pulses) then Fourier theory tells us that all frequencies would be present in equal amounts and they would be completely uncorrelated in relative phase. This is the condition we call white light. Its spatial and temporal coherence are zero (i.e. there is perfect *incoherence*) and, again, it is an unachievable fiction. Between the two fictions of perfect coherence and perfect incoherence there lies the real world.

In this real world we have to deal with real sources of light. Real sources always have a nonzero spectral width, i.e. the light power is spread over a range of frequencies, and the result of superimposing all the frequency components, as again we know from Fourier theory, is to produce a disturbance which is not a pure sine wave: an example is the wave packet we have just considered. Another is the infinitesimally narrow δ-function.

Hence there is seen to be a clear connection between spectral width and coherence. The two are inversely proportional. The narrower is a pulse of light (or any other waveform) the less it is like a pure sine wave and, by Fourier theory, the greater is its spectral width.

In the practical case of a two-slit interference pattern, the pattern will be sharp and clear if the light used is of narrow linewidth (a laser perhaps); but if a broad-linewidth source, such as a tungsten filament lamp, is used, the pattern is multicoloured, messy and confused.

Of course, we need to quantify these ideas properly if we are to use them effectively.

A2.1.4.2 Measure of coherence

If we are to determine the measurable effect which the degree of coherence is to have on optical systems, especially those which involve interference and diffraction phenomena, we need a quantitative measure of coherence. This must measure the extent to which we can, knowing the (complex) amplitude of a periodic disturbance at one place and/or time, predict its magnitude at another place and/or time. We know that this measure can be expected to have its maximum value for a pure sine wave, and its minimum value for white light. A convenient definition will render these values 1 and 0, respectively.

To fix ideas, and to simplify matters, let us consider first just temporal coherence.

We may sensibly postulate that for a time function $f(t)$ a knowledge of its value at t will provide us with some knowledge of its value up to a later time, say t', when it becomes completely independent of its value at t. If two time functions are completely independent then we expect the average value of their product over a time which is long compared with the characteristic time constant for their variations (i.e. the reciprocal of bandwidth) to be equal to the product of their individual average values, i.e.

$$\langle f(t)f(t')\rangle = \langle f(t)\rangle\langle f(t')\rangle \tag{A2.1.17}$$

and if, as is the case for the vast majority of optical disturbances, the functions oscillate about a zero mean, i.e.

$$\langle f(t)\rangle = \langle f(t')\rangle = 0$$

then it follows from equation (A2.1.17) that:

$$\langle f(t)f(t')\rangle = 0.$$

In words: the product of two independent functions, when averaged over a term which is long compared with the time over which each changes significantly, is zero.

On the other hand, if we set $t = t'$, our 'delay average' above must have its maximum possible value, since a knowledge of the value of $f(t)$ at t enables us to predict its value at that time t with absolute certainty! Hence we have

$$\langle f(t)f(t)\rangle = \langle f^2(t)\rangle$$

and this must be the maximum value of the 'delay-average' function.

Clearly, then the value of this product (for all real-world functions) will fall off from a value of unity when $t' = t$, to a value of zero when the two variations are completely independent, at some other value of t'; and it is also clear that the larger the value of t' for which this occurs, the stronger is the dependence and thus the greater is the coherence. It might well be, then, that the quantity which we seek in order to measure the coherence is a quantity which characterizes the speed at which this product function decays to zero.

Suppose, for example, we consider a pure, temporal sine wave in this context. We know in advance that this is a perfectly coherent disturbance, and conveniently we would thus require our coherence measure to be unity.

Let us write the wave as:

$$f(t) = a \sin \omega t.$$

To obtain the 'delay-average' function we first multiply this by a replica of itself, displaced by time τ (i.e. $t' = t + \tau$) (figure A2.1.21). We have:

$$\begin{aligned} f(t)f(t+\tau) &= a \sin\omega t\, a \sin\omega(t+\tau) \\ &= \frac{1}{2}a^2[\cos\omega\tau - \cos(2\omega t + \omega\tau)]. \end{aligned}$$

We now average this over all time (zero bandwidth gives a characteristic time constant of infinity!) and, since $\langle\cos(2\omega t + \omega\tau)\rangle = 0$, we have:

$$\langle f(t)f(t+\tau)\rangle = \frac{1}{2}a^2\cos\omega\tau.$$

This quantity we call the self-correlation function (sometimes the auto-correlation function), $c(\tau)$, of the disturbance. In more formal mathematical terms we would calculate it according to:

$$c(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_0^T f(t)f(t+\tau)\,\mathrm{d}t \qquad \text{(A2.1.18)}$$

i.e.

$$c(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \frac{1}{2}a^2[\cos\omega\tau - \cos(2\omega t + \tau)]t$$

and thus, again:

$$c(\tau) = \frac{1}{2}a^2\cos\omega\tau.$$

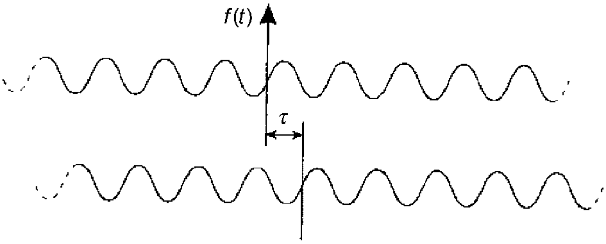


Figure A2.1.21. Self-correlation delay.

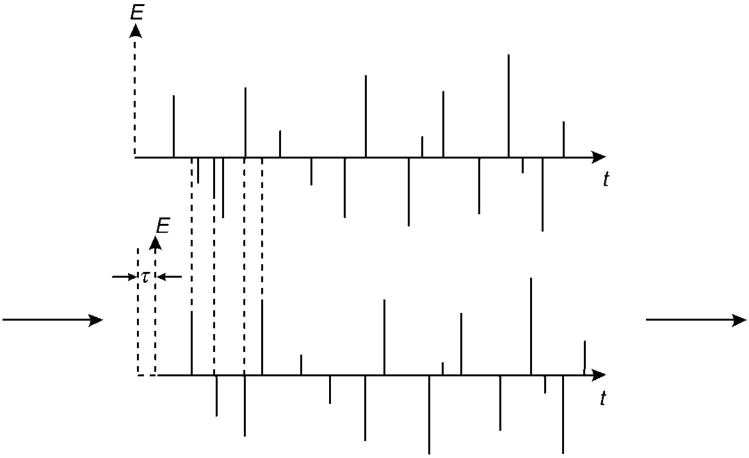


Figure A2.1.22. Auto-correlation for randomly-spaced δ-functions.

This function does not decay, but oscillates with frequency ω and constant *amplitude* $a^2/2$. It is this latter amplitude which we take as our measure of coherence of the light wave, since the sinusoidal term in $\omega\tau$ will always be present for an oscillatory field such as an optical wave, and provides no useful information on the coherence. The mathematical form of $c(\tau)$ is known as a 'convolution' integral.

Since we require, for convenience, that this measure be unity for the sine wave, we choose to normalize it to its value at $\tau = 0$ [i.e. $c(0) = a^2/2$, in this case].

This normalized function we call the coherence function $\gamma(\tau)$, and hence for the case we have considered, $\gamma(\tau) = 1$ (perfect coherence).

Consider now white light. We have noted that this is equivalent to a series of randomly spaced δ-functions which are also randomly positive or negative (figure A2.1.22). Clearly, this series must have a mean amplitude of zero if it is to represent a spread of optical sinusoids over an infinite frequency range, each with a mean amplitude of zero. If now to obtain the 'delay-average' function we multiply this set of δ-functions by a displaced replica of itself, then only a fraction of the total will overlap, and an overlap between two δ-functions of the same sign will have equal probability with that between two of opposite signs. Consequently, the mean value of the overlap function will also be zero, always, regardless of the time delay. Hence for this case, $c(\tau) = 0$ for all τ and $\gamma(\tau) = 0$, (i.e. perfect incoherence).

Consider, finally, a random stream of quanta, or wave packets (figure A2.1.23).

The packets run into each other, but each packet is largely coherent within itself. If this stream waveform is multiplied by a displaced replica of itself the result will be of the form shown in figure A2.1.24(*a*). Only when the displacement exceeds the duration of one packet does the correlation fall essentially to zero. Thus, in this case, we have a decaying sine wave, and the quantity which characterizes the decay rate of its amplitude will be our measure of coherence (for example, time to $1/e$ point for an exponential decay).

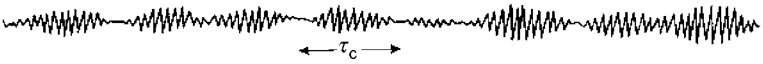


Figure A2.1.23. A random stream of wave packets.

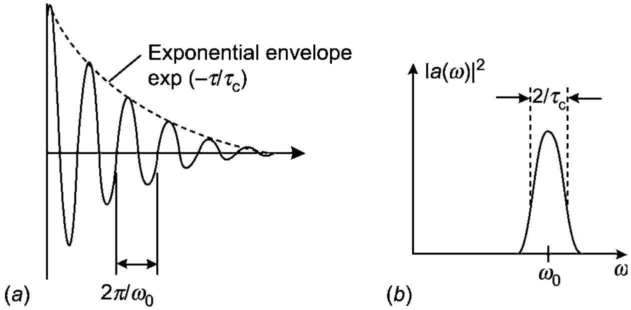


Figure A2.1.24. Stream correlation.

All of the above requirements are taken care of in the general mathematical expression for the coherence function:

$$\gamma(\tau) = \frac{\left|\int_0^\infty f(t) f^*(t+\tau)\mathrm{d}t\right|}{\int_0^\infty f(t) f^*(t)\mathrm{d}t} = \frac{|c(\tau)|}{|c(0)|}.$$

The integration performs the time averaging, the use of the complex form allows the complex conjugate in one of the functions to remove the oscillatory term in the complex exponential representation, the use of the modulus operation returns the complex value to a real value, and the division effects the required normalization. The function $c(\tau)$, as defined in equation (A2.1.18), is called the correlation coefficient and is sometimes separately useful. Note that it is, in general, a complex quantity.

To cement ideas let us just see how these functions work, again for the pure sine wave.

We must express the sine wave as a complex exponential, so that we write:

$$f(t) = a \exp(\mathrm{i}\omega t).$$

Then we have:

$$c(\tau) = \int_0^\infty a \exp(\mathrm{i}\omega t) a \exp[-\mathrm{i}\omega(t+\tau)]\,\mathrm{d}t = a^2 \exp(-\mathrm{i}\omega t)[T]_0^\infty \to \infty$$

and

$$c(0) = \int_0^\infty a \exp(\mathrm{i}\omega t)\, a \exp(-\mathrm{i}\omega t)\mathrm{d}t = a^2[T]_0^\infty \to \infty.$$

Hence

$$\gamma(\tau) = \frac{|c(\tau)|}{|c(0)|} = |\exp(\mathrm{i}\omega\tau)| = 1$$

as required.

The coherence functions for the three cases we have been considering are shown in figure A2.1.25. The sine wave's function does not decay and therefore its coherence time is infinite; the white light's function decays in zero time and thus its coherence time is zero; the 'stream of wave packets' function decays in time τ_c, and thus it is partially coherent with coherence time τ_c. All other temporal functions may have their coherence time quantified in this way. It is clear, also, that the quantity $c\tau_c$ (where c is the velocity of light), will specify a coherence length.

The same ideas, fairly obviously, can also be used for spatial coherence, with τ replaced by σ, the spatial delay. Specifically, in this case:

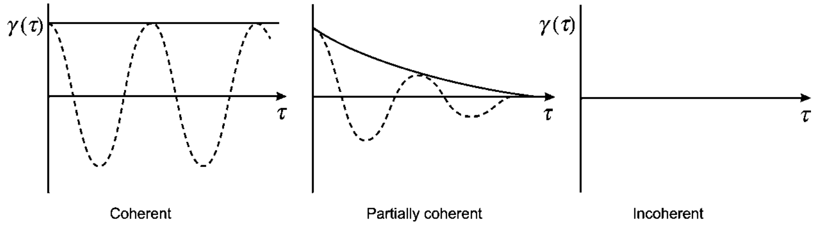


Figure A2.1.25. Coherence functions.

$$\gamma(\sigma) = \frac{\left|\int_0^\infty f(s)\, f^*(s+\sigma)\mathrm{d}s\right|}{\int_0^\infty f(s)\, f^*(s)\mathrm{d}s} = \frac{|c(\sigma)|}{|c(0)|}$$

and a ‘decay’ parameter σ_c will define the coherence length.

Finally, the mutual coherence of two separate functions $f_1(t)$ and $f_2(t)$ may be characterized by a closely similar *mutual coherence* function:

$$\gamma_{12}(\tau) = \frac{\left|\int_0^\infty f_1(t)\, f_2^*(t+\tau)\mathrm{d}t\right|}{\int_0^\infty f_1(t)\, f_2^*(t)\mathrm{d}t} = \frac{|c_{12}(\tau)|}{|c_{12}(0)|}$$

with t, τ again replaceable by s, σ, respectively, for the mutual spatial coherence case.

γ_{12} is sometimes called the ‘degree of coherence’ between the two functions.

A2.1.4.3 Dual-beam interference

We shall first consider in more detail the conditions for interference between two light beams (figure A2.1.26). It is clear, from our previous look at this topic that interference fringes will be formed if the two waves bear a constant phase relationship to each other, but we must now consider the form of the interference pattern for varying degrees of mutual coherence. In particular, we must consider the ‘visibility’ of the pattern; in other words the extent to which it contains measurable structure and contrast.

At the point O (in figure A2.1.26) the (complex) amplitude resulting from the two sources P_1 and P_2 is given by:

$$A = f_1(t'') + f_2(t'')$$

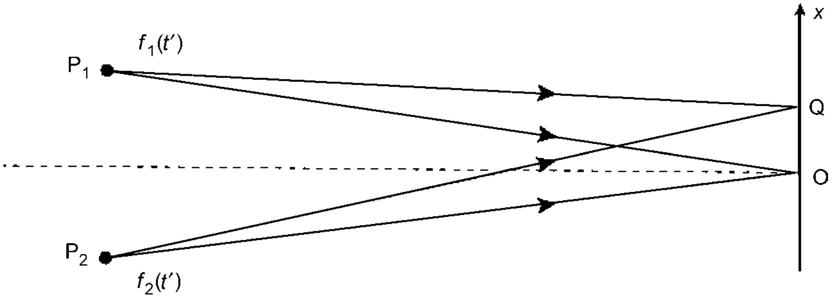


Figure A2.1.26. Two-source interference.

where $t'' = t' + \tau_0$, τ_0 being the time taken for light to travel from P_1 or P_2 to O. If f_1, f_2 represent the electric field amplitudes of the waves, the observed intensity at O will be given by the square of the modulus of this complex number. Remember that the modulus of the complex number: $A = a + \mathrm{i}b$ is written

$$|a + \mathrm{i}b| = (a^2 + b^2)^{1/2}$$

and a convenient stratagem for obtaining the square of the modulus of a complex number is to multiply it by its complex conjugate, i.e.:

$$AA^* = (a + \mathrm{i}b)(a - \mathrm{i}b) = (a^2 + b^2).$$

Hence, in this case the optical intensity is given by

$$I_0 = \langle AA^* \rangle = \langle [f_1(t'') + f_2(t'')][f_1^*(t'') + f_2^*(t'')] \rangle$$

where the triangular brackets indicate an average taken over the response time of the detector (e.g. the human eye) and we assume that f_1 and f_2 contain the required constant of proportionality ($K^{1/2}$) to relate optical intensity with electric field strength, i.e. $I = KE^2$.

At point Q the amplitudes will be:

$$f_1\left(t'' - \frac{1}{2}\tau\right), \quad f_2\left(t'' + \frac{1}{2}\tau\right)$$

τ being the time difference between paths P_2Q and P_1Q.

Writing $t = t' - \frac{1}{2}\tau$ we have the intensity at Q:

$$I_Q = \langle [f_1(t) + f_2(t + \tau)][f_1^*(t) + f_2^*(t + \tau)] \rangle$$

i.e.

$$I_Q = \langle f_1(t)f_1^*(t) \rangle + \langle f_2(t)f_2^*(t) \rangle + \langle f_2(t + \tau)f_1^*(t) \rangle + \langle f_1(t)f_2^*(t + \tau) \rangle.$$

The first two terms are clearly the independent intensities of the two sources at Q. The second two terms have the form of our previously defined mutual correlation function, in fact:

$$\langle f_1(t)f_2^*(t + \tau) \rangle = c_{12}(\tau) \quad \langle f_1^*(t)f_2(t + \tau) \rangle = c_{12}^*(\tau).$$

We may note, in passing, that each of these terms will be zero if f_1 and f_2 have orthogonal polarizations, since in that case neither field amplitude has a component in the direction of the other, there can be no superposition, and the two cannot interfere. Hence the average value of their product is again just the product of their averages, each of which is zero, being a sinusoid.

If $c_{12}(t)$ is now written in the form:

$$c_{12}(\tau) = |c_{12}(\tau)| \exp(\mathrm{i}\omega\tau)$$

(which is valid provided that f_1 and f_2 are sinusoids in ωt) we have:

$$c_{12}(\tau) + c_{12}^*(\tau) = 2|c_{12}(\tau)| \cos \omega\tau.$$

Hence, provided that we observe the light intensity at Q with a detector which has a response time very much greater than the coherence times (self and mutual) of the sources (so that the time averages are valid), then we may write the intensity at Q as:

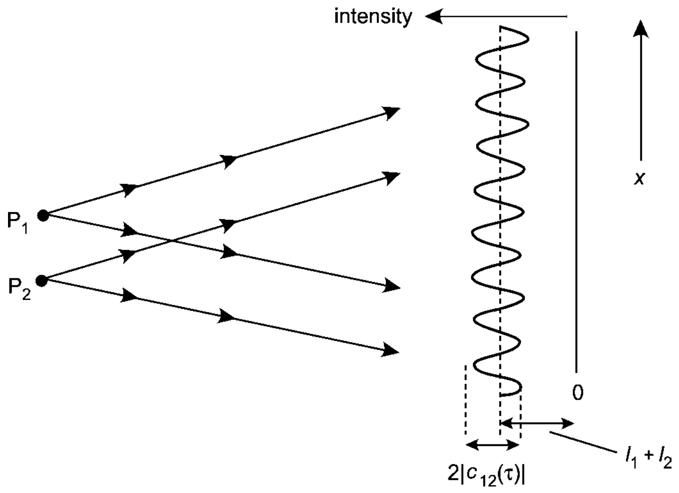


Figure A2.1.27. Mutual coherence function ($|c_{12}(\tau)|$) from the two-source interference pattern.

$$I_Q = I_1 + I_2 + 2|c_{12}(\tau)| \cos \omega t. \tag{A2.1.19}$$

As we move along x we shall effectively increase τ, so we shall see a variation in intensity whose amplitude will be $2|c_{12}(\tau)|$ (i.e. twice the modulus of the mutual coherence function) and which varies about a mean value equal to the sum of the two intensities (figure A2.1.27). Thus we have an experimental method by which the mutual coherence of the sources, $c_{12}(t)$, can be measured.

If we now define a fringe *visibility* for this interference pattern by:

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$

which quantifies the contrast in the pattern, i.e. the difference between maxima and minima as a fraction of the mean level, then, from equation (A2.1.19):

$$V(\tau) = \frac{2|c_{12}(\tau)|}{(I_1 + I_2)}$$

and with, as previously defined:

$$\gamma(\tau) = \frac{|c_{12}(\tau)|}{|c_{12}(0)|}$$

we note that:

$$|c_{12}(0)| = \left| \int_0^\infty f_1(t)\, f_2^*(t) \mathrm{d}t \right| = K\langle E_1 \rangle \langle E_2 \rangle = (I_1 I_2)^{1/2}$$

and thus:

$$\gamma\,(\tau) = \frac{|c_{12}(\tau)|}{(I_1 I_2)^{1/2}}.$$

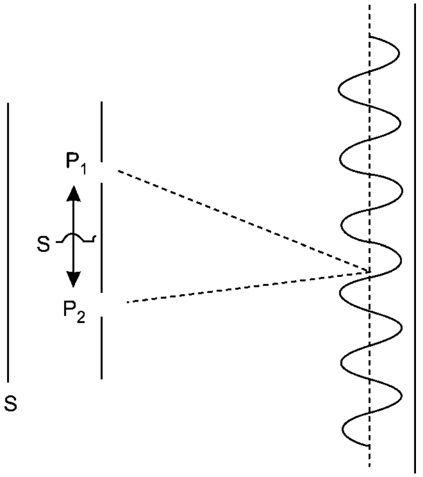


Figure A2.1.28. Extended-source interference.

Hence the visibility function $V(\tau)$ is related to the coherence function $\gamma(\tau)$ by:

$$V(\tau) = \frac{2(I_1 I_2)^{1/2}}{(I_1 + I_2)} \gamma(\tau)$$

and if the two intensities are equal, we have:

$$V(\tau) = \gamma(\tau)$$

i.e. the visibility and coherence functions are identical.

From this we may conclude that, for equal-intensity coherent sources, the visibility is 100% ($\gamma = 1$); for incoherent sources it is zero; and for partially coherent sources the visibility gives a direct measure of the actual coherence.

If we arrange that the points P_1 and P_2 are pinholes equidistant from and illuminated by a single source S, then the visibility function clearly measures the self-coherence of S. Moreover, the Fourier transform of the function will yield the source's power spectrum.

Suppose now that the two holes are placed in front of an extended source, S, as shown in figure A2.1.28, and that their separation is variable. The interference pattern produced by these sources of light now measures the correlation between the two corresponding points on the extended source. If the separation is initially zero and is increased until the visibility first falls to zero, the value of the separation at which this occurs defines a spatial coherence dimension for the extended source. And if the source is isotropic, a coherence area is correspondingly defined. In other words, in this case any given source point has no phase correlation with any point which lies outside the circular area of which it is the centre point.

A2.1.4.4 Summary

In this section, we have looked at the conditions necessary for optical waves to interfere in a consistent and measurable way, with themselves and with other waves. We have seen that the conditions relate to the extent to which properties such as amplitude, phase, frequency and polarization remain constant in time and space or, to put it another way, the extent to which knowledge of the properties at one point in time or space tells us about these properties at other points.

Any interference pattern will only remain detectable as long as coherence persists and, by studying the rise and fall of interference patterns, much can be learned about the sources themselves and about the processes which act upon the light from them.

Coherence also relates critically to the information-carrying capacity of light and to our ability to control and manipulate it sensibly. The design and performance of any device or system which relies on interference or diffraction phenomena must take into account the coherence properties of the sources which are to be used; some of these work to the designer's disadvantage, but others do not.

A2.1.5 Optical waveguiding

A2.1.5.1 Introduction

The basic principles of optical waveguiding are quite straightforward. Waves are guided when they are constrained to lie within a channel between two other media, the refractive index of the channel material being slightly higher than those of the other media, so that the light can 'bounce' along the channel by means of a series of total internal reflections (TIRs) at the boundaries between media. The case shown in figure A2.1.29 is that where a channel of refractive index n_1 lies between two slabs, each with refractive index n_2 ($n_1 > n_2$); this is the easiest arrangement to analyse mathematically, yet it illustrates all the important principles.

The other important point is that, in order to progress down the guide indefinitely, the waves from the successive boundaries must interfere constructively, forming what is essentially a continuous, stable, interference pattern down the guiding channel. If the interference is not fully constructive, the wave will eventually 'self-destruct', owing to the out-of-phase cancellations (although, clearly, if the phasings are almost correct, the wave might persist for a considerable distance, attenuating only slowly). The condition which must be imposed for constructive interference defines for us the guided wave parameters, in particular those angles of bounce which can give rise to the 'modes' of the waveguide, i.e. the various patterns of constructive interference which are allowed by the restrictions (boundary conditions) of the guide geometry.

The ideas involved in waveguiding are thus quite simple. In order to make use of them we need, as always, a proper mathematical description, so we shall in this section develop this description.

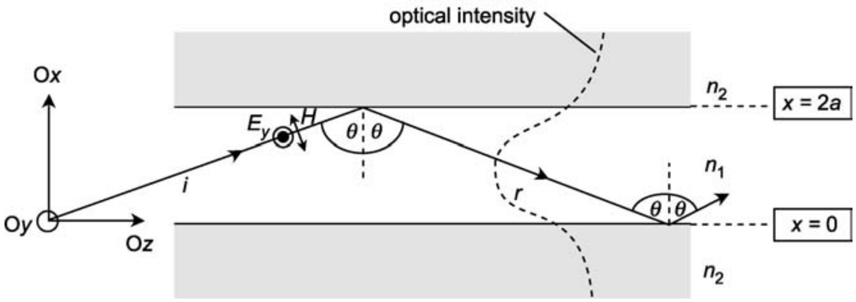


Figure A2.1.29. Optical slab waveguide.

A2.1.5.2 The planar waveguide

We begin by considering the symmetrical slab waveguide shown in figure A2.1.29. The guiding channel consists here of a slab of material of refractive index n_1 surrounded by two outer slabs, each of refractive index n_2. The resultant electric field for light which is linearly polarized in a direction perpendicular to the plane of incidence (the so-called transverse electronic (TE) mode) is given by the sum of the upwards and downwards propagating rays:

$$E_{\mathrm{T}} = E_{\mathrm{i}} + E_{\mathrm{r}}$$

where

$$E_{\mathrm{i}} = E_0 \exp(\mathrm{i}\omega t - \mathrm{i}kn_1 x \cos\vartheta - \mathrm{i}kn_1 z \sin\vartheta)$$

(i.e. a wave travelling in the xz plane at an angle ϑ to the slab boundaries which lie parallel to the yz plane) and

$$E_{\mathrm{r}} = E_0 \exp(\mathrm{i}\omega t + \mathrm{i}kn_1 x \cos\vartheta - \mathrm{i}kn_1 z \sin\vartheta + \mathrm{i}\delta_{\mathrm{s}})$$

which is the wave resulting from the reflection at the boundary, and differs from E_{i} in two aspects: it is now travelling in the negative direction of Ox, hence the change of sign in the x term, and there has been a change of phase at the reflection, hence the $\mathrm{i}\delta_{\mathrm{s}}$ term. We must also remember that δ_{s} depends on the angle, ϑ, the polarization of the wave and, of course, n_1 and n_2 according to the Fresnel equation [4b]. Hence

$$E_{\mathrm{T}} = E_{\mathrm{i}} + E_{\mathrm{r}} \quad = 2E_0 \cos\left(kn_1 x \cos\vartheta + \frac{1}{2}\delta_{\mathrm{s}}\right) \exp\left(\mathrm{i}\omega t - \mathrm{i}kn_1 z \sin\vartheta + \mathrm{i}\frac{1}{2}\delta_{\mathrm{s}}\right) \qquad (\mathrm{A2.1.20}a)$$

which is a wave propagating in the Oz direction, but with amplitude varying in the Ox direction according to $2E_0 \cos\left(kn_1 x \cos\vartheta + \frac{1}{2}\delta_{\mathrm{s}}\right)$ (see figure A2.1.29).

The symmetry of the arrangement tells us that the intensity ($\sim$ *square* of the electric field) of the wave must be the same at the two boundaries and thus that it is the same at $x = 0$ as at $x = 2a$. Hence:

$$\cos^2\left(\frac{1}{2}\delta_{\mathrm{s}}\right) = \cos^2\left(kn_1 2a \cos\vartheta + \frac{1}{2}\delta_{\mathrm{s}}\right)$$

which implies that:

$$2akn_1 \cos\vartheta + \delta_{\mathrm{s}} = m\pi \qquad (\mathrm{A2.1.20}b)$$

where m is an integer. This is our 'transverse resonance condition' and it is a condition on ϑ (remember that δ_{s} also depends on ϑ) which defines a number of allowed values for ϑ (corresponding to the various integer values of m) which in turn define our discrete, allowable modes (or interference patterns) of propagation.

Now the wavenumber, $k = 2\pi/\lambda$, for the free space propagation of the wave has suffered a number of modifications. First, the wavelength of the light is smaller in the medium than in free space (the frequency remains the same, but the velocity is reduced by a factor $n_{1,2}$), so we can conveniently define

$$\beta_1 = n_1 k \quad \beta_2 = n_2 k$$

as the wavenumbers in the guiding and outer slabs, respectively. Secondly, however, if we choose to interpret equation (A2.1.20a) as one describing a wave propagating in the Oz direction with amplitude modulated in the Ox direction, it is convenient to resolve the wavenumber in the guiding medium into components along Oz and Ox, i.e.

along Oz:

$$\beta = n_1 k \sin\vartheta \qquad (\mathrm{A2.1.21}a)$$

along *Ox*

$$q = n_1 k \cos\vartheta. \tag{A2.1.21$b$}$$

Of these two components β is clearly the more important, since it is the effective wavenumber for the propagation down the guide. In fact, equation (A2.1.20*a*) can now be written:

$$E_\mathrm{T} = 2E_0 \cos\left(qx + \frac{1}{2}\delta_\mathrm{s}\right) \exp \mathrm{i}\left(\omega t - \beta z + \mathrm{i}\frac{1}{2}\delta_\mathrm{s}\right).$$

What can be said about the velocity of the wave down the guide? Clearly the phase velocity is given by:

$$c_\mathrm{p} = \frac{\omega}{\beta}.$$

However, we know that this is not the end of the story, for the velocity with which optical energy propagates down the guide is given by the group velocity [5], which, in this case is given by

$$c_\mathrm{g} = \frac{\mathrm{d}\omega}{\mathrm{d}\beta}.$$

What, then, is the dependence of ω upon β?

To answer this, let us start with equation (A2.1.21*a*), i.e.

$$\beta = n_1 k \sin\vartheta.$$

The first thing to note is that, for all real ϑ, this requires:

$$\beta \le n_1 k.$$

Also, since the TIR condition requires that

$$\sin\vartheta \ge \frac{n_2}{n_1}$$

if follows that:

$$\beta = n_1 k \sin\vartheta \ge n_2 k.$$

Hence we have

$$n_1 k \ge \beta \ge n_2 k$$

or

$$\beta_1 \ge \beta \ge \beta_2.$$

In other words, the wavenumber describing the propagation along the guide axis always lies between the wavenumbers for the guiding medium (β_1) and the outer medium (β_2). This we might have expected from the physics, since the propagation lies partly in the guide and partly in the outer medium (evanescent wave). We shall be returning to this point later.

Remember that our present concern is about how β varies with ω between these two limits, so how else does equation (A2.1.21*a*) help?

Clearly, the relation

$$k = \frac{\omega}{c_0}$$

where c_0 is the free space velocity, gives one dependence of β on ω, but what about $\sin\vartheta$? For a given value of m (i.e. a given mode) the transverse resonant condition (A2.1.20*b*) provides the dependence of ϑ on k. However, this is quite complex since, as we know, δ_s is a quite complex function of ϑ. Hence in order to proceed further this dependence must be considered.

The expressions for the phase changes which occur under TIR at a given angle are well known [4]:

$$\tan\frac{1}{2}\delta_s = \frac{(n_1^2\sin^2\vartheta - n_2^2)^{1/2}}{n_1\cos\vartheta}$$

for the case where the electric field is perpendicular to the plane of incidence and

$$\tan\frac{1}{2}\delta_p = \frac{n_1(n_1^2\sin^2\vartheta - n_2^2)^{1/2}}{n_2^2\cos\vartheta}$$

for the case where it lies in the plane of incidence.

Note also that:

$$\tan\frac{1}{2}\delta_p = \frac{n_1^2}{n_2^2}\tan\frac{1}{2}\delta_s.$$

Finally, let us define, for convenience, a parameter, p, where

$$p^2 = \beta^2 - n_2^2k^2 = k^2(n_1^2\sin^2\vartheta - n_2^2). \qquad \text{(A2.1.22)}$$

The physical significance of p will soon become clear.

We now discover that we can cast our ‘transverse resonance’ condition (A2.1.20*b*) into the form

$$\tan\left(aq - \frac{1}{2}m\pi\right) = \frac{p}{q} \quad (E_\perp) \qquad \text{(A2.1.23}a\text{)}$$

for the perpendicular polarization and

$$\tan\left(aq - \frac{1}{2}m\pi\right) = \frac{n_1^2}{n_2^2}\frac{p}{q} \quad (E_\parallel) \qquad \text{(A2.1.23}b\text{)}$$

for the parallel polarization.

The conventional waveguide notation designates these two cases as ‘transverse electric (TE)’ for $E_\perp$ and ‘transverse magnetic (TM)’ for $E_\parallel$. The terms refer, of course, to the direction of the stated fields with respect to the plane of incidence of the ray.

We can use equations (A2.1.23*a*) and (A2.1.23*b*) to characterize the modes for any given slab geometry. The solutions of the equations can be separated into odd and even types according to whether m is odd or even. For odd m we have

$$\tan\left(aq - \frac{1}{2}m_{\mathrm{odd}}\pi\right) = \cot aq \qquad \text{(A2.1.24}a\text{)}$$

and for even m

$$\tan\left(aq - \frac{1}{2}m_{\mathrm{even}}\pi\right) = \tan aq. \qquad \text{(A2.1.24}b\text{)}$$

Taking m to be even we may then write equation (A2.1.23a), for example, in the form:

$$aq \tan aq = ap \quad (E_\perp). \tag{A2.1.25}$$

Now from the definitions of p and q it is clear that:

$$a^2p^2 + a^2q^2 = a^2k^2(n_1^2 - n_2^2). \tag{A2.1.26}$$

Taking rectangular axes ap, aq this latter relation between p and q translates into a circle of radius $ak(n_1^2 - n_2^2)^{1/2}$ (figure A2.1.30). If, on the same axes, we also plot the function aq tan aq, then equation (A2.1.25) is satisfied at all points of intersection between the two functions (figure A2.1.30). (A similar set of solutions clearly can be found for odd m.) These points, therefore, provide the values of ϑ which correspond to the allowed modes of the guide. Having determined a value for ϑ for a given k, β can be determined from:

$$\beta = n_1 k \sin \vartheta$$

and hence β can be determined as a function of k (for a given m) for the TE modes.

Now, finally, with

$$k = \frac{\omega}{c}$$

we have the relationship between β and ω which we have been seeking. For obvious reasons these are called dispersion curves, and are important determinants of waveguide behaviour. They are drawn either

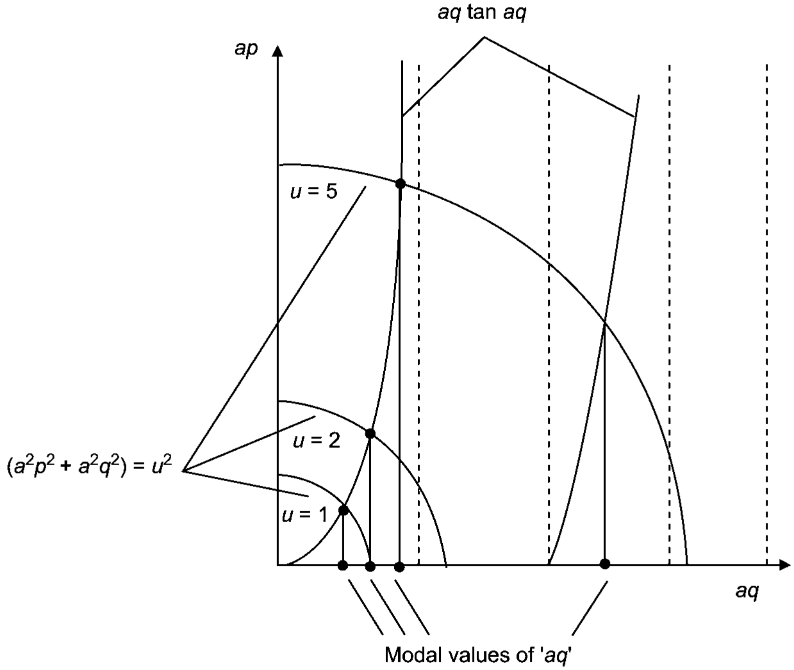


Figure A2.1.30. Graphical solution of the modal equation for the slab waveguide.

as β versus k or as ω versus β. The three lowest order modes for a typical slab waveguide are shown in figure A2.1.31(*a*) using the latter representation. Clearly, this is the more convenient form for determining the group velocity $\mathrm{d}\omega/\mathrm{d}\beta$ by simple differentiation (figure A2.1.31(*b*)).

A final point of great importance should be made. As k decreases, so the quantity

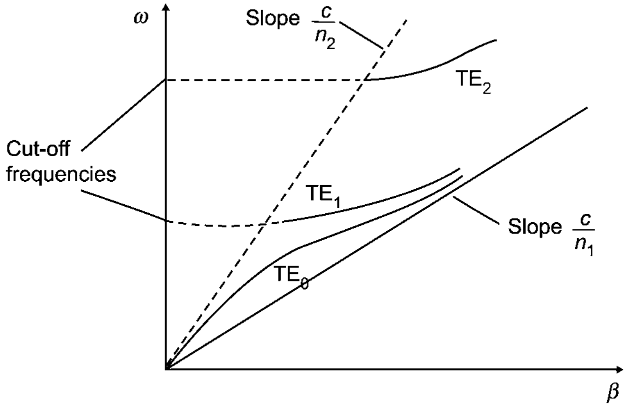
$$a^2p^2 + a^2q^2 = a^2k^2(n_1^2 - n_2^2)$$

decreases and the various modes are sequentially 'cut-off' as the circle (figure A2.1.30) reduces in radius. This is also apparent in figure A2.1.31(*a*) since a reduction in k corresponds to a reduction in ω. Clearly the number of possible modes depends upon the waveguide parameters a, n_1 and n_2. However, it is also clear that there will always be at least one mode, since the circle will always intersect the tan curve at one point, even for a vanishingly small circle radius. If there is only one solution, then figure A2.1.30 shows that the radius of the circle must be less than $\pi/2$, i.e.

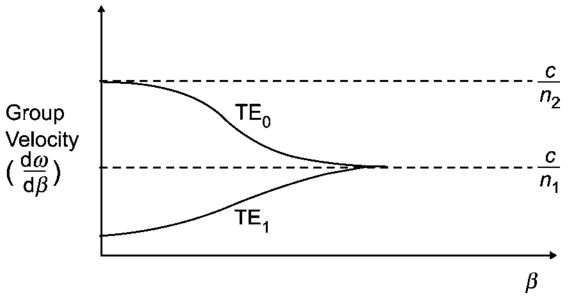
$$ak(n_1^2 - n_2^2)^{1/2} < \frac{1}{2}\pi$$

or

$$\frac{2\pi a}{\lambda}(n_1^2 - n_2^2)^{1/2} < 1.57. \tag{A2.1.27}$$



(*a*) Dispersion diagram for slab waveguide



(*b*) Variation of group velocity with wavenumber

Figure A2.1.31. Dispersion and group velocity for a slab waveguide.

This quantity is another important waveguide parameter, for this and many other reasons. It is given the symbol V and is called the 'normalized frequency', or, quite often, simply the 'V number'. Thus,

$$V = \frac{2\pi a}{\lambda}(n_1^2 - n_2^2)^{1/2}.$$

Equation (A2.1.27) is thus the single-mode condition for this symmetrical slab waveguide. It represents an important case, since the existence of just one mode in a waveguide simplifies considerably the behaviour of radiation within it, and thus facilitates its use in, for example, the transmission of information along it. Physically, equation (A2.1.27) is stating the condition under which it is impossible for constructive interference to occur for any ray other than that which (almost) travels along the guide axis.

Clearly, a very similar analysis can be performed for the TM modes, using equation (A2.1.23*b*).

Look again now at figure A2.1.29. It is clear that there are waves travelling in the outer media with amplitudes falling off the farther we go from the central channel. This is a direct result of the necessity for fields (and their derivatives) to be continuous across the media boundaries. We know from equation (A2.1.20*a*) that the field amplitude in the central channel varies as:

$$E_x = 2E_0 \cos\left(kn_1 x \cos\vartheta + \frac{1}{2}\delta_s\right).$$

How does the field in the outer slabs vary? The evanescent field in the second medium, when TIR occurs, falls off in amplitude according to

$$E_x = E_a \exp\left(-\frac{2\pi x}{\lambda_2}\sinh\gamma\right); \quad x > a$$

where:

(i) E_a is the value of the field at the boundary, i.e.

$$E_a = 2E_0 \cos\left(kn_1 a \cos\vartheta + \frac{1}{2}\delta_s\right)$$

(ii) λ_2 is the wavelength in the second medium, and is equal to λ/n_2

(iii) $2\pi\sinh\gamma/\lambda_2 = k(n_1^2\sin^2\vartheta - n_2^2)^{1/2}$ and this can now be identified with p from equation (A2.1.22). Hence:

$$E_x = E_a \exp(-px); \quad x > a$$

and we see that p is just the exponential decay constant for the amplitude of the evanescent wave (figure A2.1.32) and, from equation (A2.1.22), we note that $p \sim 0.1k$. (It is a fact of any physical analysis that all parameters of mathematical importance will always have a simple physical meaning.)

So the evanescent waves are waves which propagate in the outer media parallel with the boundary but with amplitude falling off exponentially with distance from the boundary.

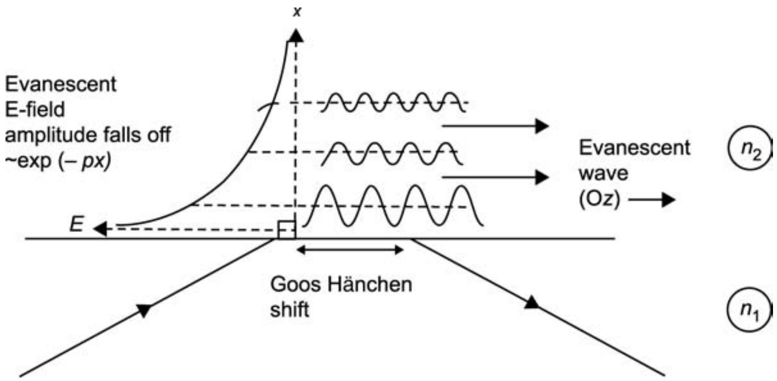


Figure A2.1.32. Evanescent wave decay.

These evanescent waves are very important. First, if the total propagation is not to be disturbed the thickness of each outer slab must be great enough for the evanescent wave to have negligible amplitude at its outer boundary: the wave falls off as $\sim\exp(-x/\lambda)$, so at $x \sim 20\lambda$ it normally will be quite negligible ($\sim 10^{-9}$). At optical wavelengths, then, the slabs should have a thickness $\geq 20\,\mu$m.

Secondly, since energy is travelling (in the *Oz* direction!) in the outer media, the evanescent wave properties will influence the core propagation, in respect, for example, of loss and dispersion.

A2.1.5.3 Integrated optics

Planar waveguides find interesting application in integrated optics. In this, waves are guided by planar channels and are processed in a variety of ways. An example is shown in figure A2.1.33. This is an electro-optic modulator and it utilises electro-optic effect (see section A2.1.7.7) whereby the application of an electric field to a medium alters its refractive index. However, the electric field is acting on a waveguide which, in this case, is a channel (such as we have just been considering) surrounded by ‘outer slabs’ called here a ‘substrate’. The electric field is imposed by means of the two substrate electrodes, and the interaction path is under close control, as a result of the waveguiding. The material of which both the substrate and the waveguide are made should, in this case, clearly be an electro-optic material, such as

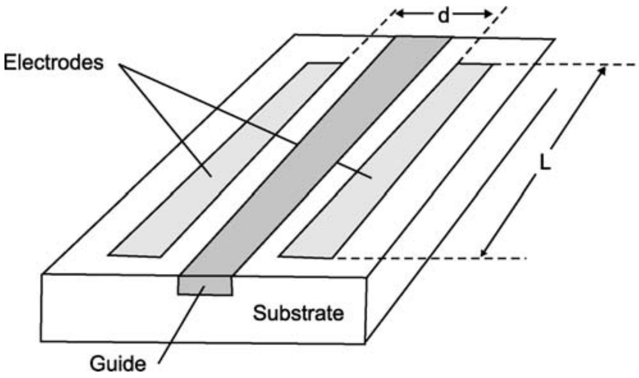


Figure A2.1.33. An integrated-optical electro-optic phase modulator.

lithium tantalate ($LiTaO_3$) whose refractive index depends upon the applied electric field. The central waveguiding channel may be constructed by diffusing ions into it (under careful control); an example of a suitable ion is niobium (Nb), which will thus increase the refractive index of the 'diffused' region and allow total internal reflection to occur at its boundaries with the 'raw' $LiTaO_3$. Many other functions are possible using suitable materials, geometries and field influences. It is possible to fabricate couplers, amplifiers, polarizers, filters, etc. all within a planar 'integrated' geometry.

One of the advantages of this integrated optics technology is that the structures can be produced to high manufactured tolerance by 'mass production' methods, allowing them to be produced cheaply if required numbers are large, as is likely to be the case in optical telecommunications, for example.

A fairly recent, but potentially very powerful, development is that of the 'optoelectronic integrated circuit' (OEIC) which combines optical waveguide functions with electronic ones such as optical source control, photodetection and signal processing, again on a single, planar, readily-manufacturable 'chip'.

Note, finally, that in figure A2.1.33 the 'upper' slab (air) has a different refractive index from the lower one (substrate). This is thus an example of an asymmetrical planar waveguide, the analysis of which is more complex than the symmetrical one which we have considered. However, the basic principles are the same; the mathematics is just more cumbersome, and is covered in many other texts [6].

A2.1.5.4 Cylindrical waveguides

Let us now consider the cylindrical dielectric structure shown in figure A2.1.34. This is the geometry of the optical fibre, the central region being known as the 'core' and the outer region as the 'cladding'. In this case the same basic principles apply as for the dielectric slab, but the circular, rather than planar, symmetry complicates the mathematics. We use, for convenience, cylindrical co-ordinates (r, ϑ, z) as defined in figure A2.1.34. This allows us to cast Maxwell's wave equation for the dielectric structure into the form:

$$\nabla^2 E = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial E}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 E}{\partial \varphi^2} + \frac{\partial^2 E}{\partial z^2} = \mu\varepsilon\frac{\partial^2 E}{\partial t^2}. \tag{A2.1.28}$$

If we try a solution for E in which all variables are separable, we write:

$$E = E_r(r)E_\varphi(\varphi)E_z(z)E_t(t)$$

and can immediately, from the known physics, take it that:

$$E_z(z)E_t(t) = \exp[\mathrm{i}(\beta z - \omega t)].$$

In other words the wave is progressing along the axis of the cylinder with wavenumber β and with angular frequency ω. It follows, of course, that its (phase) velocity of progression along the axis is given by:

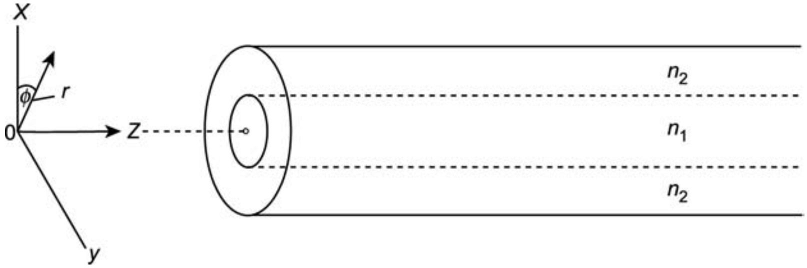


Figure A2.1.34. Cylindrical waveguide geometry.

$$c_{\mathrm{p}} = \frac{\omega}{\beta}.$$

By substitution of these expressions into the wave equation (A2.1.28) we may rewrite it in the form:

$$\frac{\partial}{\partial r}\left(r\frac{\partial(E_r E_\varphi)}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2(E_r E_\varphi)}{\partial\varphi^2} - \beta^2 E_r E_\varphi + \mu\varepsilon\omega^2 E_r E_\varphi = 0.$$

Now if we suggest a periodic function for E_ϑ of the form:

$$E_\varphi = \exp(\pm \mathrm{i} l\varphi)$$

where l is an integer, we can further reduce the equation to:

$$\frac{\partial^2 E_r}{\partial r^2} + \frac{1}{r}\frac{\partial E_r}{\partial r} + \left(n^2k^2 - \beta^2 - \frac{l^2}{r^2}\right)E_r = 0.$$

This is a form of Bessel's equation, and its solutions are Bessel functions (see any advanced mathematical text, e.g. [7]). If we use the same substitutions as for the previous planar case, i.e.

$$n_1^2k^2 - \beta^2 = q^2$$
$$\beta^2 - n_2^2k^2 = p^2$$

we find for $r \le a$ (core)

$$\frac{\partial^2 E_r}{\partial r^2} + \frac{1}{r}\frac{\partial E_r}{\partial r} + \left(q^2 - \frac{l^2}{r^2}\right)E_r = 0$$

and for $r > a$ (cladding)

$$\frac{\partial^2 E_r}{\partial r^2} + \frac{1}{r}\frac{\partial E_r}{\partial r} + \left(p^2 + \frac{l^2}{r^2}\right)E_r = 0.$$

Solutions of these equations are (see figure A2.1.35(*a*)):

$$E_r = E_{\mathrm{c}} J_l(qr);\quad r \le a$$
$$E_r = E_{\mathrm{cl}} K_l(pr);\quad r > a$$

where J_l is a 'Bessel function of the first kind' and K_l is a 'modified Bessel function of the second kind' (sometimes known as a 'modified Hankel function'). The two functions must clearly be continuous at $r = a$, and we have for our full 'trial' solution in the core

$$E = E_{\mathrm{c}} J_l(qr)\exp(\pm \mathrm{i} l\varphi)\exp\ \mathrm{i}(\beta z - \omega t)$$

and a similar one for the cladding

$$E = E_{\mathrm{cl}} J_l(pr)\exp(\pm \mathrm{i} l\varphi)\exp\ \mathrm{i}(\beta z - \omega t).$$

Again we can determine the allowable values for p, q and β by imposing the boundary conditions at $r = a$ [8]. The result is a relationship which provides the β versus k, or 'dispersion' curves, shown in figure A2.1.36. The mathematical manipulations are tedious, but are somewhat eased by using the

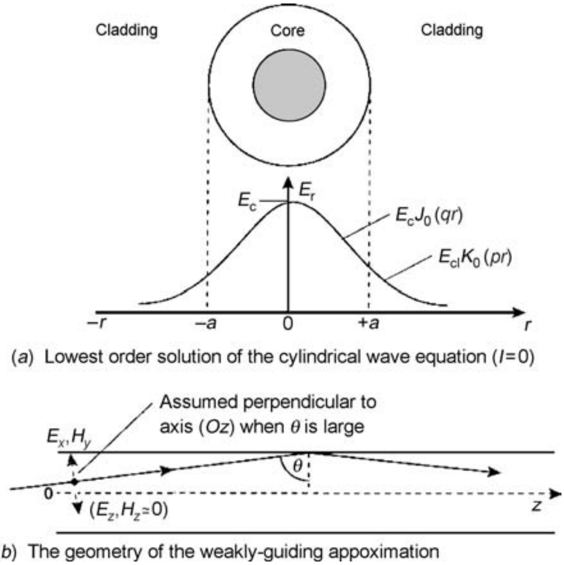

Cladding
Core
Cladding
E_c
E_r
$E_cJ_0(qr)$
$E_{cl}K_0(pr)$
$-r$
$-a$
0
$+a$
r
(a) Lowest order solution of the cylindrical wave equation ($l=0$)
Assumed perpendicular to axis (Oz) when θ is large
E_x, H_y
θ
0
$(E_z, H_z = 0)$
z
(b) The geometry of the weakly-guiding appoximation

Figure A2.1.35. Solution for the cylindrical waveguide equation, and the weakly-guiding approximation.

so-called 'weakly guiding' approximation. This makes use of the fact that if $n_1 \sim n_2$, then the ray's angle of incidence on the boundary must be very large, if TIR is to occur. The ray must bounce down the core almost at grazing incidence. This means that the wave is very nearly a transverse wave, with very small z components. By neglecting the longitudinal components H_z, E_z, a considerable simplification of the mathematics results (figure A2.1.35(b)). Since the wave is, to a first approximation, transverse, it can be resolved conveniently into two linearly polarized components, just as for free space propagation. The

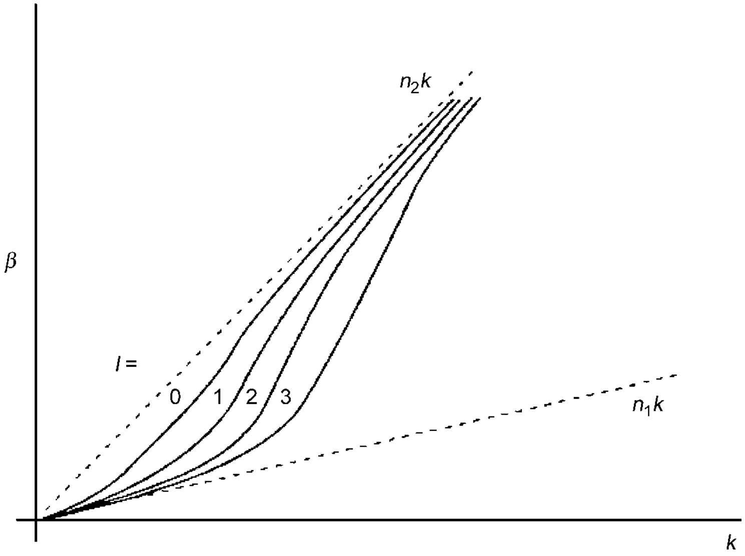


Figure A2.1.36. Dispersion curves for the cylindrical waveguide.

modes are thus dubbed ‘linearly polarized (LP)’ modes, and the notation which describes the profile’s intensity distribution is the ‘LP’ notation.

A2.1.5.5 Optical fibres

The cylindrical geometry relates directly to the optical fibre. The latter has just the geometry we have been considering and, for a typical fibre:

$$\frac{n_1 - n_2}{n_1} \sim 0.01$$

so that the weakly guiding approximation is valid. Some of the low-order ‘LP modes’ of intensity distribution are shown in figure A2.1.37, together with their polarizations, and values for the azimuthal integer, 1. There are, then, two possible linearly polarized optical fibre modes. For the cylindrical geometry the ‘single-mode condition’ is (analogously to equation (A2.1.24) for the planar case):

$$V = \frac{2\pi a}{\lambda}(n_1^2 - n_2^2)^{1/2} < 2.405.$$

The figure 2.405 derives from the value of the argument for which the lowest order Bessel function, J_0, has its first zero. Some important practical features of optical-fibre design can be appreciated by reversion to geometrical (ray) optics.

Let us consider, first, the problem of launching light into the fibre. Referring to figure A2.1.38(*a*), we have for a ray incident on the front face of the fibre at angle ϑ_0, and with refracted angle ϑ_1:

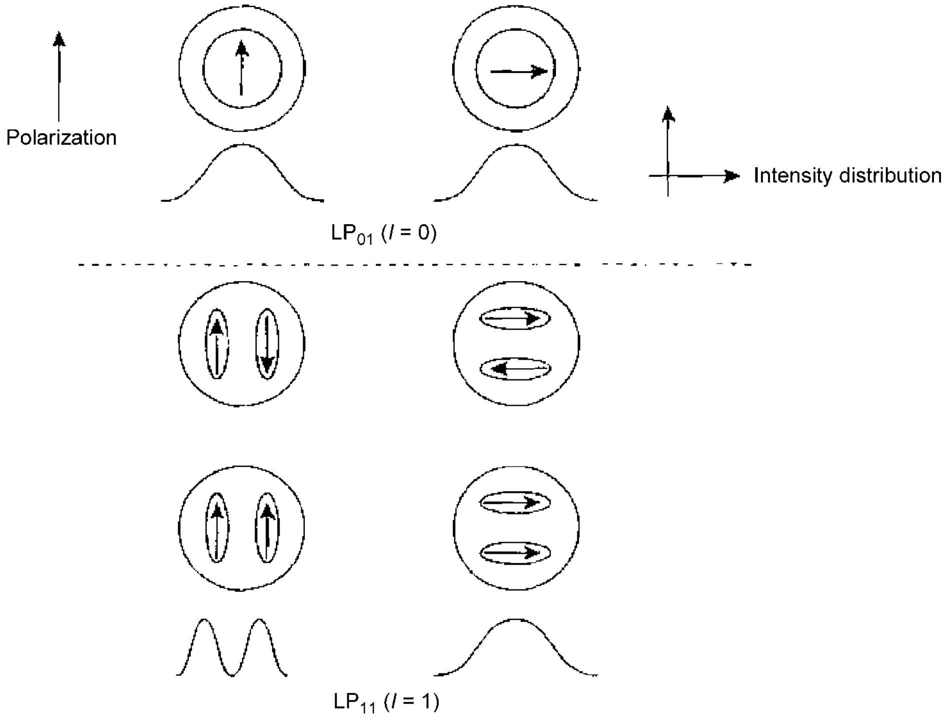
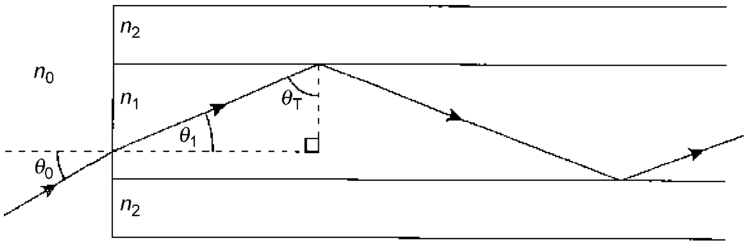
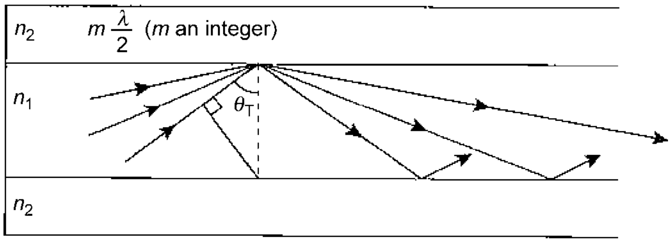


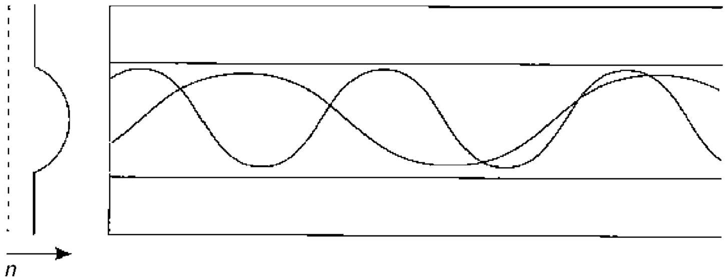
Figure A2.1.37. Some low order modes of the cylindrical waveguide (with weakly-guiding mode labels).


n_2
n_0
n_1
θ_T
θ_1
θ_0
n_2

(*a*) Acceptance angle for an optical fibre


n_2
$m\frac{\lambda}{2}$ (m an integer)
n_1
θ_T
n_2

(*b*) Ray representations of fibre modes



(*c*) Graded index ray paths

Figure A2.1.38. Ray propagation in optical fibres.

$$n_0 \sin\vartheta_0 = n_1 \sin\vartheta_1$$

where n_0 and n_1 are the refractive indices of air and the fibre core material, respectively. If the angle at which the ray then strikes the core/cladding boundary is ϑ_T, then, for TIR, we must have: $\sin\vartheta > n_2/n_1$ where n_2 is the cladding index.

Since $\vartheta_T = (1/2)\pi - \vartheta_1$ the inequality is equivalent to

$$\cos\vartheta_1 > \frac{n_2}{n_1}$$

so from Snell's law expression above,

$$\cos\vartheta_1 = \left(1 - \frac{n_0^2 \sin^2\vartheta_0}{n_1^2}\right)^{1/2}$$

or

$$n_0 \sin \vartheta_0 < (n_1^2 - n_2^2)^{1/2}.$$

The quantity on the RHS of this inequality is known as the numerical aperture (NA) of the fibre. It is a specification of the 'acceptance' cone of light, this being a cone of apex half-angle ϑ_0. Clearly, a large refractive index difference between core and cladding is necessary for a large acceptance angle; for a typical fibre, $\vartheta_0 \sim 10°$.

The discrete values of reflection angle which are allowed by the transverse resonance condition (within the TIR condition) can be represented by the ray propagations shown in figure A2.1.38(*b*). This makes clear that for a large number of allowable rays (i.e. modes) the TIR angle should be large, implying a large NA. However it is also clear, geometrically, that the rays will progress down the guide at velocities which depend on their angles of reflection: the smaller the angle, the smaller the velocity. This leads to large 'modal dispersion' at large NA since, if the launched light energy is distributed among many modes, the differing velocities will lead to varying times of arrival of the energy components at the far end of the fibre. This is undesirable in, for example, communications applications, since it will lead to a limitation on the communications bandwidth. In a digital system, a pulse cannot be allowed to spread into the pulses before or after it. For greatest bandwidth only one mode should be allowed, and this requires a small NA. Thus a balance must be struck between good signal level (large NA) and large signal bandwidth (small NA).

A fibre design which attempts to attain a better-balanced position between these is shown in figure A2.1.38(*c*). This fibre is known as graded-index (GI) fibre and it possesses a core refractive index profile which falls off parabolically (approximately) from its peak value on the axis. This profile constitutes, effectively, a continuous convex lens, which allows large acceptance angle while limiting the number of allowable modes to a relatively small value. GI fibre is used widely in short and medium distance communications systems. For trunk systems single-mode fibre is invariably used, however. This ensures that the modal dispersion is entirely absent, thus removing this limitation on bandwidth. Single-mode fibre possesses a communications bandwidth which is an order of magnitude greater than that of multimode fibre.

A2.1.5.6 Summary

Optical waveguiding is of primary importance to the optoelectronic designer. With its aid it is possible to confine light and to direct it to where it is needed, over short, medium and long distances.

Furthermore, with the advantage of confinement, it is possible to control the interaction of light with other influences, such as electric, magnetic or acoustic fields, which may be needed to impress information upon it. Control also can be exerted over its intensity distribution, its polarization state and its nonlinear behaviour; this last topic is covered in section A2.1.7.

In short, optical waveguiding is crucial to the control of light. For the designers of devices and systems (especially telecommunications systems, using optical fibres) this control is essential.

A2.1.6 Electrons in solids

A2.1.6.1 Introduction

In order to understand the mechanisms involved in the operation of important solid state devices such as semi-conductor lasers, light-emitting diodes, various types of photodetectors, light modulators, etc it is necessary to look into some of the rather special features of the behaviour of electrons in solid materials, and this is the subject of the present section.

A solid is a state of matter where the constituent atoms or molecules are held in a rigid structure as a result of the fact that the inter-molecular forces are large compared with the forces of thermal motion of the molecules. This can only be true if the molecules are close together, for the molecules are electronically neutral overall, and forces can only exist between them if there is significant overlap among the wavefunctions of the outer electrons. This overlap leads to another important consequence: the energy levels in which the electrons lie are shared levels; they are a property of the material as a whole rather than of the individual molecules, as is the case for a gas, for example.

In order to gain a physical 'feel' for the effect of the strong interaction on the energy level structure in a solid, consider what happens when two simple oscillators, such as two pendulums, interact. If two pendulums each of same length, and thus with the same independent frequency of oscillation, f, are strung from the same support bar, they will interact with each other via the stresses transmitted through the bar, as they swing. For the combined system there are two 'eigenmodes', that is to say two states of oscillation which are stable in time. These are the state where the two pendulums swing together, in phase, and that where they swing in opposition, in antiphase. For all the other states the amplitudes and relative phases of the two pendulums vary with time. The two eigenstates have difference frequencies f_p (in phase) and f_a (in anti-phase) and we find that

$$f_p > f > f_a.$$

The original frequency f is not now a characterizing parameter of the system, having been replaced by two other frequencies, one higher and one lower. If just one of the two pendulums is set swinging it will do so at a frequency in the range f_p to f_a and will set the other pendulum swinging. The second pendulum will acquire maximum amplitude when the first has come to a stop and then the process will reverse. The energy will continuously transfer between the pendulums at a frequency $(f_p - f_a)$. Consider now *three* identical pendulums strung from the same bar. Now there are three eigenstates: (i) all in phase; (ii) outer two in phase, central one in antiphase; (iii) left- or right-hand two in phase, right- or left-hand one in antiphase. Each of these states has its own frequency of oscillation, so we now have three frequencies. It is an easy conceptual extrapolation to n pendulums, where there will be n frequencies centred on the original f, i.e. the original single frequency has become a band of n frequencies. If n is very large, as it is with the number of molecules in a solid, the frequencies are so close together as to comprise essentially a continuous *band* of frequencies, and thus also of electron energy levels. Thus we can expect each discrete energy level of the isolated atom or molecule to form a separate band of allowable energies, and the bands will be separated by gaps which represent energies forbidden to electrons (figure A2.1.39). This feature is crucial to the behaviour of electrons in solids and accounts for most of the properties which are important in optoelectronics. It is, therefore, necessary to study it in more detail before looking at why, exactly, it is so important to us.

A2.1.6.2 Elements of the band theory of solids

Having understood why energy bands form in solids, it is necessary now to understand the ways in which electrons occupy them. This is as crucial to the understanding of the optoelectronic properties of solids as the formation of the bands themselves.

First, it is necessary to remember that electrons obey quantum rules. Associated with each electron is a wave whose wavenumber (~reciprocal of wavelength) is related to the momentum (p) of the electron.

If the electron is propagating freely the relationship is:

$$p = \frac{h}{\lambda} = \frac{hk}{2\pi}$$

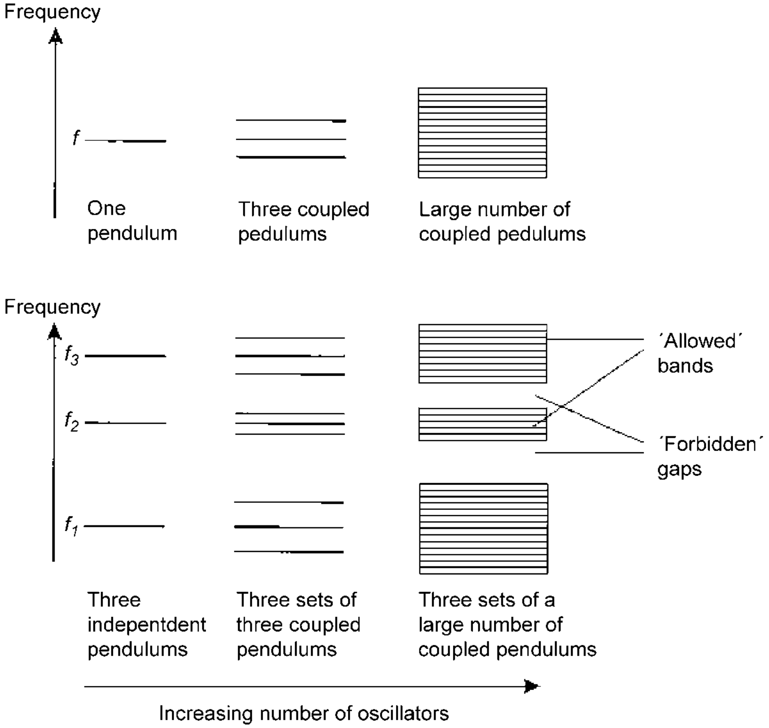

Frequency
f
One pendulum
Three coupled pedulums
Large number of coupled pedulums
Frequency
f_3
f_2
f_1
'Allowed' bands
'Forbidden' gaps
Three indepentdent pendulums
Three sets of three coupled pendulums
Three sets of a large number of coupled pendulums
Increasing number of oscillators

Figure A2.1.39. Band structure resulting from the coupling of oscillators.

and since the kinetic energy of a particle of mass m is related to its momentum p by:

$$E_{\mathrm{k}} = \frac{p^2}{2m}.$$

We have, in the case of the free electron,

$$E_{\mathrm{k}} = \frac{h^2k^2}{8\pi^2 m}. \tag{A2.1.29}$$

We shall need this shortly.

Another consequence of the quantum behaviour of electrons is that they distribute themselves among available energy levels in a rather special way. We say that their distribution obeys 'Fermi–Dirac' statistics, and although it is not necessary to go very deeply into this, it is necessary to understand the basic ideas.

All fundamental particles, such as electrons, protons, neutrons, mesons, quarks, etc, are indistinguishable particles, i.e. there is no way in which an electron, say, can be 'labelled' at one time or place, in such a way that it is possible to recognize it as the same particle at another time or place; and this is not just a 'labelling' problem: it is quite fundamental—a consequence of quantum physics. Hence if two identical (indistinguishable) particles are interchanged in any energy distribution within a system there can be no change in any of the observable macroscopic properties of the system. Now these observable properties depend only on the square of the modulus of the system's overall wavefunction (section A1.8), which is formed from all of the individual electron wavefunctions, i.e.

$$\psi = \psi(1)\psi(2)\cdots\psi(n).$$

If electrons 1 and 2 are interchanged then $|\psi|^2$ must remain the same, i.e.

$$|\psi_{12}|^2 = |\psi_{21}|^2$$

hence

$$\psi_{12} = \pm\psi_{21}.$$

This presents two possibilities: either the interchange leaves the sign of the wavefunction the same, or it reverses it. Particles which leave the sign the same are called symmetrical particles; particles which reverse it are called anti-symmetrical particles.

Now comes the vital point: two anti-symmetrical particles cannot occupy the same quantum state, since the interchange of two identical particles occupying the same quantum state cannot alter the wavefunction in any way at all, not even its sign. Hence no two anti-symmetrical particles can have the same set of 'quantum numbers', numbers which define the quantum state uniquely. This is known as the Pauli exclusion principle. Electrons are anti-symmetrical particles and thus obey the Pauli exclusion principle. In fact all particles with 'half-integral spin', $(n + \frac{1}{2})h/2\pi$, obey the principle, e.g. electrons, protons, neutrons, μ-mesons; these are called fermions (note the small *f* now!). Particles with integral spin, $nh/2\pi$, are symmetrical particles and obey 'Bose–Einstein' statistics: e.g. photons, α-particles, π-mesons; these are called bosons.

The fact that no two electrons can occupy the same quantum state is profound, and is the single most important feature of the behaviour of electrons, in regard to the optoelectronic properties of solids. It means that the available electrons will fill the available quantum states progressively and systematically from bottom to top, like balls in a vertical tube whose diameter is just sufficient to take one ball at a time.

Let us examine this 'filling' process in more detail.

Each allowed energy level in any system contains (in general) more than one quantum state. The number of states which it contains is called the 'degeneracy' of the energy level. (Remember also that each of the bands in the solid state energy structure results from a large number of closely spaced energy levels, so there is also a kind of multiple degeneracy within a band.)

Now suppose, firstly, that the electrons within a given energy band are completely free to move around as if they were an electron 'gas' in the solid. This is approximately true for electrons in a metal, and the only restriction really is that the electrons are not free to leave the solid. How should we calculate the energy states available to the electrons in this case?

Well, fortunately, most of the work necessary for calculating the number of electron states which lie between energies E and $E + \mathrm{d}E$ for this case has already been done, in section A2.1.2.1, for atomic oscillators which give rise to electromagnetic waves; the analogy between electromagnetic waves in a box and electrons in a box is very close. The electron waves are restricted to the same set of discrete values by the box boundaries as were the electromagnetic waves. The only difference is that whereas we had to allow for two polarization states in the electromagnetic case, we now have to allow for two spin directions (e.g. up and down) in the electron case. In both cases we must multiply by a factor of 2, so that equation (A2.1.6*a*) remains valid; i.e. the number of electron states with k values between k and $k + \mathrm{d}k$ is $g(k)$ where

$$g(k)\mathrm{d}k = \frac{k^2\mathrm{d}k}{\pi^2}.$$

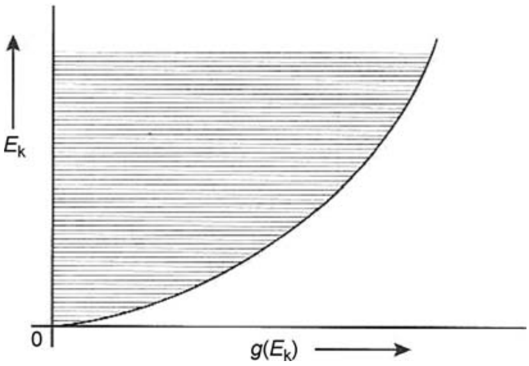


Figure A2.1.40. 'Density of states' function for a metal.

$g(k)$ is known as the degeneracy function. All that is necessary now is to express this in terms of the energy by substituting for k and dk from equation (A2.1.29), i.e.:

$$g(E_{\mathrm{k}})\mathrm{d}E_{\mathrm{k}} = \frac{1}{\pi^2}\frac{8\pi^2 m}{h^2}E_{\mathrm{k}}\left(\frac{8\pi^2 m}{h^2}\right)^{1/2}\frac{1}{2}E_k^{-1/2}\mathrm{d}E_{\mathrm{k}}$$

$$= \frac{4\pi}{h^3}(2m)^{3/2}E_k^{1/2}\mathrm{d}E_{\mathrm{k}}. \qquad \text{(A2.1.30}a\text{)}$$

This function is shown in figure A2.1.40, and in solid state parlance, is usually called the 'density of states' function. It represents the number of states between energies E_{k} and $E_{\mathrm{k}} + \mathrm{d}E_{\mathrm{k}}$.

Hence these are the states which are going to be filled from the bottom up. Each range of E_{k} to $E_{\mathrm{k}} + \mathrm{d}E_{\mathrm{k}}$ will be filled sequentially like the balls in the tube. What, then, is the occupancy of these states? How do the electrons actually distribute themselves among them?

If each energy range is filled in turn, the best way to express this is to plot the fraction of the $g(E_{\mathrm{k}})$ levels which is filled by a total of N_{T} electrons. At the absolute zero of temperature this occupancy function will look like variation A in figure A2.1.41. All the states will be filled up to the level at which the electrons are exhausted. Hence up to that level the fractional occupancy is 1; above that level, it is zero. This level is known as the Fermi level, E_{F}, and is easily calculated if the total number of electrons, N_{T}, is known, for it is necessary only to integrate equation (A2.1.30a) between 0 and E_{F}:

$$N_{\mathrm{T}} = \frac{4\pi}{h^3}(2m)^{3/2}\int_0^{E_{\mathrm{F}}} E_{\mathrm{k}}^{1/2}\mathrm{d}E_{\mathrm{k}}.$$

Hence

$$E_{\mathrm{F}} = \left(\frac{3N_{\mathrm{T}}}{8\pi}\right)^{2/3}\frac{h^2}{2m}. \qquad \text{(A2.1.31)}$$

Using equation (A2.1.31) the density of states function (A2.1.30a) may now conveniently be expressed in the form:

$$g(E)\mathrm{d}E = \frac{3}{2}N_{\mathrm{T}}\frac{E^{1/2}}{E_{\mathrm{F}}^{3/2}}\mathrm{d}E. \qquad \text{(A2.1.30}b\text{)}$$

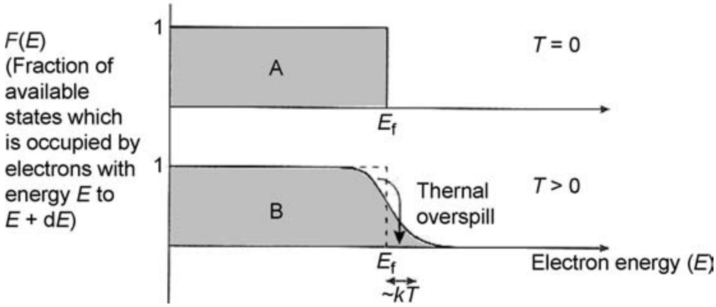


Figure A2.1.41. Fermi–Dirac 'occupancy' distributions.

Suppose now that the temperature rises above absolute zero to a small value $T(>0)$. This makes available to each electron an extra energy $\sim kT(\ll E_F)$. However, most of the electrons cannot take advantage of this because the next available empty level for them is more than kT away. The only electrons which can gain energy are those at the top of the distribution, for they have empty states above them. The distribution thus changes to the one shown as B in figure A2.1.41 for temperature T. Hence the electrons behave very differently from a gas, say, where the average energy of *all* the molecules would increase by kT. (It is for this reason that the specific heat of metals is much smaller ($\sim 1\%$) than was predicted on a free electron theory of metallic conduction; this discrepancy was a great puzzle to physicists in the early years of this century.)

The function which describes the occupancy of the levels at a given temperature is known as the Fermi–Dirac function. It is given by:

$$F(E) = \frac{1}{\exp\left(\frac{E - E_F}{kT}\right) + 1}. \tag{A2.1.32a}$$

Note that it has the behaviour which has already been described:

For $T = 0$ and

$$E < E_F : \exp\left(\frac{E - E_F}{kT}\right) \to 0, \quad F(E) \to 1.$$

For $T = 0$ and

$$E > E_F : \exp\left(\frac{E - E_F}{kT}\right) \to \infty, \quad F(E) \to 0.$$

This clearly corresponds to variation A in figure A2.1.41.

As the temperature rises, the topmost electrons move to higher states and the function develops a 'tail', whose width is $\sim kT$ (curve B in figure A2.1.41). The energy E_F in this case corresponds to the energy for which $F(E) = 0.5$.

Now we are in a position to make the final step: the density of electrons within a given small energy range will be the product of the density of quantum states and the actual occupancy of these states. It will be the product of the density of states function (A2.1.30*b*) and the Fermi–Dirac function (A2.1.32), i.e. $n(E)\mathrm{d}E = g(E)F(E)\mathrm{d}E$ or

$$n(E)\mathrm{d}E = \frac{3}{2}N_{\mathrm{T}}\frac{E^{1/2}}{E_{\mathrm{F}}^{3/2}}\frac{\mathrm{d}E}{\exp\left(\dfrac{E-E_{\mathrm{F}}}{kT}\right)+1} \qquad \text{(A2.1.33}a\text{)}$$

where $n(E)\mathrm{d}E$ is the number per unit volume of electrons with energies between E and $E + \mathrm{d}E$. This function is shown in figure A2.1.42 for $T = 0$ and for $T \neq 0$.

It is interesting to note, before leaving this, that the Fermi–Dirac distribution is a prevalent feature primarily because, in a solid, the number of electrons is comparable with the number of quantum states, and therefore the electrons must be carefully packed according to the quantum rules. If the number of quantum states far exceeds the number of identical particles, as it does in a gas for example, the quantum rules are scarcely noticeable. To see this suppose that, in equation (A2.1.30*b*), $g(E) \gg N_{\mathrm{T}}$, then $E^{1/2} \gg E_{\mathrm{F}}^{3/2}$ and hence $E \gg E_{\mathrm{F}}$. Equation (A2.1.33*a*) becomes:

$$n(E)\mathrm{d}E = \frac{3}{2}N_T\frac{E^{1/2}}{E_{\mathrm{F}}^{3/2}}\exp\left(-\frac{E}{kT}\right)\mathrm{d}E. \qquad \text{(A2.1.33}b\text{)}$$

Expressed in terms of molecular velocity, v, and remembering that the molecular energy in this case is purely kinetic energy of motion, i.e.

$$E = \frac{1}{2}mv^2$$

we have:

$$n(v)\mathrm{d}v = Av^2\exp\left(-\frac{mv^2}{2kT}\right)\mathrm{d}v$$

which is the Maxwell–Boltzmann gas velocity distribution as deduced from classical (i.e. nonquantum) statistical thermodynamics [1].

It turns out that equation (A2.1.33*b*) often also represents a useful approximation in solid state physics. In all cases where the electron distribution is being considered well above the Fermi level (i.e. $E \gg E_{\mathrm{F}}$) the Fermi–Dirac distribution function of equation (A2.1.32*a*) approximates to:

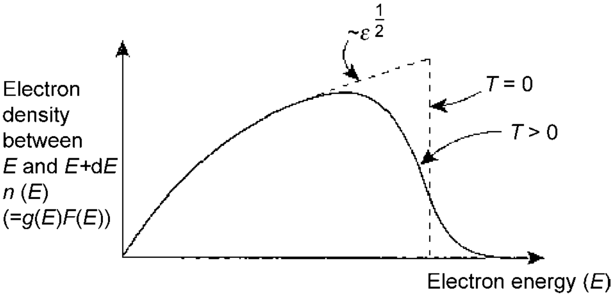


Figure A2.1.42. Electron density distributions for a metal.

$$F(E) = \exp\left(-\frac{E}{kT}\right) \tag{A2.1.32b}$$

which is, of course, the Boltzmann factor. We shall have several occasions to use this later.

A2.1.6.3 Metals, insulators and semiconductors

We are now in a position to understand, qualitatively at first, what it is that distinguishes metals, insulators and semiconductors. It all depends upon the position of the Fermi level.

Consider the solid state band structure in figure A2.1.43. Suppose, first, that a solid consists of atoms or molecules with just one electron in the outermost energy shell. This shell forms a band of energy levels in the solid, as we have seen, and the total number of available states will be $2N$ per unit volume, where N is the number of atoms per unit volume (i.e. two electron spin states per quantum state). But there will be only N electrons per unit volume since there is only one electron per atom. Hence the band is half-filled, and the Fermi level lies halfway up the second band, as in figure A2.1.43(*a*). The electrons at the top of the Fermi–Dirac distribution have easy access to the quantum states above them and can thus move freely in response to, for example, an applied electric field, by gaining small amounts of kinetic energy from it; they can also move to conduct heat energy quickly and easily: we have a metal.

Suppose, secondly, that there are two electrons in the outer shell of the atoms or molecules comprising the solid. The band formed from the shell is now just full and the Fermi level is above the top of the band, as in figure A2.1.43(*b*). The electrons at the top of the band now can only increase their energies by jumping up to the next band. If the energy gap is quite large, neither moderate temperatures nor moderate electric fields can provide sufficient energy for this to happen. Hence, the material does not conduct electricity at all easily: we have an insulator.

Finally, consider the case shown in figure A2.1.43(*c*), again a case where the uppermost level is just full (which will, clearly, be the case for any even number of electrons in the outer shell). Here the Fermi level lies about halfway up the gap between the valence and conduction bands and the gap is now relatively small, say less than $100kT$ for room temperature. (For example the element silicon has a band gap of 1.1 eV, compared with a value for kT, at room temperature, of $\sim 2.5 \times 10^{-2}$ eV. An electron would gain an energy of 1 eV in falling through a potential difference of 1 V.) In this case, although at low temperature the Fermi–Dirac 'tail' does not extend into the upper, conduction band, at higher temperatures it does, giving a small number of electrons in the conduction band. These electrons can then move easily into the abundance of empty states now available to them in this band. Thus the room

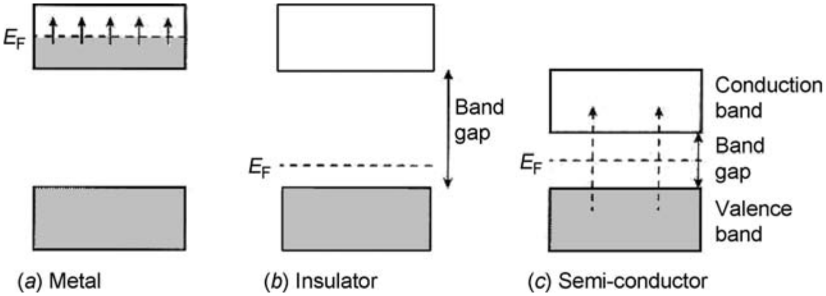


Figure A2.1.43. The Fermi level and the classification of solids.

temperature electrical conductivity is low but measurable; furthermore, it is clear that it will increase quite rapidly with temperature, as the 'tail' extends. We have here a 'semiconductor', more precisely an *intrinsic* semiconductor (it will become clear later why this adjective is necessary). For obvious reasons the upper band is called the conduction band and the lower one the valence band (since it is the stability of electrons in the lower band which provides the atomic forces holding the solid together). There is another important point to be made for the intrinsic semiconductor. When thermal agitation raises an electron from the valence band to the conduction band, it leaves behind an empty state. This state can be filled by another electron, in the valence band, which can then gain energy and contribute to the electrical conduction. These empty states, comprising, as they do, the absence of negative electric charge, are equivalent to positive 'holes' in the valence band, and they effectively move like positive charges as the electrons in the valence band move in the opposite direction to fill them. Positive holes in the valence band comprise an important feature of semiconductor behaviour, and we shall be returning to them shortly.

Before moving on it should be emphasized that the description above is a greatly simplified one, in order to establish the ideas. Solids are complicated states of matter and are three-dimensional, so in general we must not deal just with a single Fermi level but with a three-dimensional Fermi surface, which will have a shape dependent upon the variation of the material's properties with direction. Many important properties of solids depend upon the particular shape which this surface assumes. Especially important is the fact that two energy bands can sometimes overlap, so that it is possible for some elements to behave as metals even though each of their atoms possesses an even number of electrons (the lower band feeds electrons into the middle of the upper band); examples are beryllium, magnesium and zinc. However, this is the stuff of pure solid state physics and, for more, interested readers must refer to one of the many specialist texts on solid state physics [9].

It has become clear then that the position of the Fermi level in relation to the band structure for a particular solid material is vitally important. It is important not only for distinguishing between metals, insulators and semiconductors but also for understanding the detailed behaviour of any particular material.

We have seen how to calculate the Fermi level for the case of electrons moving freely within a solid. However, electrons are not entirely free to move within a crystal, or even a quasi-crystal structure, and this has important consequences which lie within a more detailed consideration of this topic. Space limitations do not allow this treatment here and the interested reader is referred to the literature [9].

A2.1.6.4 Extrinsic semiconductors

Finally, we must consider another very important type of semiconductor material. This is the doped semiconductor, otherwise known as the 'extrinsic' semiconductor. In these materials, the semiconducting properties can be both enhanced and controlled by adding specific impurities in carefully judged quantities. The effect of this is to alter the electron energy distribution in a controlled way.

We begin by considering a particular intrinsic semiconductor, silicon, since, with germanium, it is one of the two most commonly used materials for doping in this way. Both materials have a diamond-like structure, with each atom surrounded symmetrically by four others. Silicon is tetravalent, having an even number of electrons in its valence shell. There will thus be $4N$ available electrons per unit volume. The first valence energy band will be filled with $2N$ electrons and the second band also with $2N$ electrons; thus both lower bands are full and the next higher one is empty, at absolute zero (figure A2.1.44). The gap between the upper valence band and the conduction band is quite small, only 1.1 eV, compared with

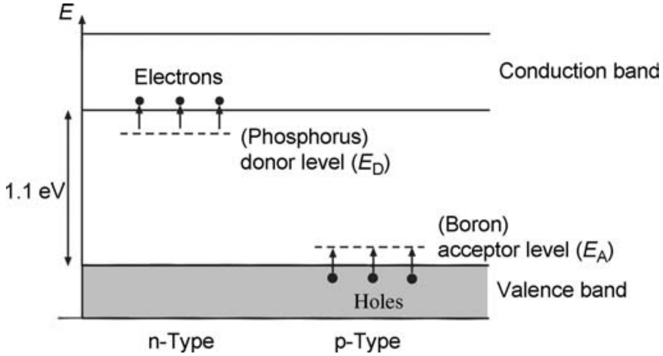


Figure A2.1.44. Energy level diagram for doped silicon.

a room temperature value of kT of $\sim 2.5 \times 10^{-2}$ eV so, although silicon is an intrinsic semiconductor, its semiconductivity is moderate, and it increases exponentially with temperature.

Suppose now that the silicon is doped with a small fraction (between 1 atom in 10^6 and 1 in 10^9) of a pentavalent (valency of five) impurity atom such as phosphorus. This atom sits quite comfortably in the silicon lattice but has one extra electron compared with the silicon atoms by which it is surrounded, and this electron is easily detached to roam through the lattice and add to the conductivity. In energy level parlance we say that it needs little energy to raise it into the conduction band; in fact for this particular case it needs only 4.5×10^{-2} eV, equivalent to only about $2kT$ at room temperature. So the energy level structure looks like figure A2.1.44 with the 'donor' level, E_D, just below the conduction band. Note that the level remains sharp, since the dopant atoms are scarce and spaced well apart in the lattice, and thus their wavefunctions cannot overlap to form a band structure, as do the atoms of the host lattice. Since, in this case, the effect of the phosphorus is to donate electrons (negative charges) to the conductivity, this is called an n-type semiconductor. The really important point is that the conductivity now is entirely under control, via control of the doping level. The greater the concentration of the dopant atoms, the greater will be the concentration of electrons in the conduction band.

Consider, on the other hand, what happens if we dope the silicon with a tervalent (valency of three) impurity such as boron. In this case the impurity atom has one electron fewer than the surrounding silicon atoms, and electrons from the valence band in the silicon can easily move into the space so created. These 'absent' (from the valence band) electrons create positive holes, as we have seen, and these also are effective in increasing the conductivity. For obvious reasons this is now called a p-type semiconductor (figure A2.1.44) and the corresponding energy level is an 'acceptor' level. The 'majority carriers' are holes in this case; in an n-type material the majority carriers are electrons.

Usually, the donor or acceptor dopings dominate the semiconductor behaviour. In other words, it is normally the case that the dopant concentrations exceed the intrinsic carrier concentration n_i. If the dopant concentrations are N_d for donor and N_a for acceptor, it must be that for charge neutrality of the material:

$$n + N_a = p + N_d. \tag{A2.1.33}$$

However, it is also true that, for all circumstances:

$$pn = n_i^2.$$

Hence, for an acceptor doping (p-type material) we have: $N_a \gg N_d, n_i$ and thus from equation (A2.1.33):

$$p = N_a$$

$$n = \frac{n_\mathrm{i}^2}{N_a}.$$

For a donor material (n-type): $N_\mathrm{d} \gg N_\mathrm{a}$, n_i and thus:

$$n = N_\mathrm{d}$$

$$p = \frac{n_\mathrm{i}^2}{N_\mathrm{d}}.$$

It is clear, then, that a knowledge of n_i and the dopant level fixes the carrier concentrations and thus allows the main features of behaviour to be determined. Where are the Fermi levels in these extrinsic semiconductors? We know that in the case of intrinsic semiconductors, the Fermi level lies about halfway between the valence and conduction bands. In an n-type semiconductor the valence band is almost full and most of the conduction is due to electrons donated from the donor levels. Hence it follows that the '50% electron occupancy' level, i.e. the Fermi level, will now lie about half way between the donor level and the bottom of the conduction band, since the top of the 'valence' band can now be identified with the donor level (figure A2.1.45).

Similarly, for p-type semiconductors, it will lie midway between the top of the valence band and the acceptor energy level. However, this can only be the case as long as the donor or acceptor mechanisms dominate. At higher temperatures most of the donor and acceptor sites would have been exhausted and the true valence band then starts to dominate the conduction mechanism. Hence the Fermi level will vary with temperature as shown in figure A2.1.45, until at high enough temperatures, it reverts to the intrinsic value in both p and n cases.

A2.1.6.5 Binary and ternary semiconductors

We cannot leave semiconductors without a mention of some important, relatively new, materials. These are alloys, made from two or more elements in roughly equal proportions and from different groups in the periodic table, and thus with differing numbers of electrons in the outermost shell.

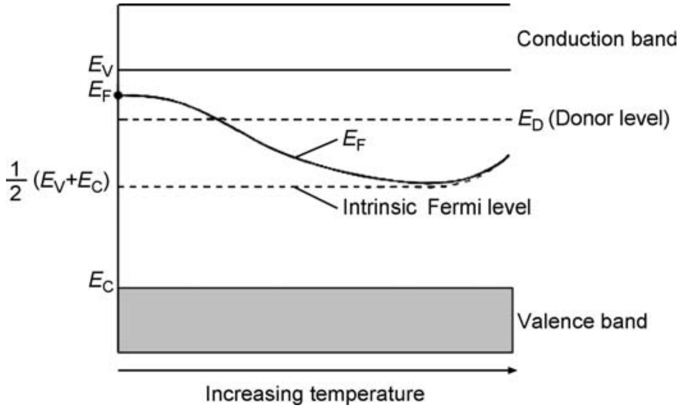


Figure A2.1.45. Fermi level in an n-type intrinsic semiconductor, and its variation with temperature.

The best known of these is gallium arsenide (GaAs) which, since Ga has a valency of III and As of V, is an example of what is called a III–V compound.

An important aspect of these compounds is that we can 'tailor' the band gap by varying the mix. The eight electrons in the two outer shells are shared to some extent and create some ionic bonding, through absence from the parent atom (i.e. it creates a positive ion). The band gap of GaAs when the two elements are present in equal proportions (i.e. same number of molecules per unit volume) is 1.4 eV, but this can be varied by replacing the As by P (GaP, 2.25 eV) or Sb (GaSb, 0.7 eV) for example. Furthermore, the materials can be made p- or n-type by increasing the V(As) over the III(Ga) component, or vice versa.

Another very important aspect of GaAs is that it is a direct-band-gap material: the minimum energy in the conduction-band Brillouin zone occurs at the same *k* value as the maximum energy of the valence band (figure A2.1.46(*a*)). This means that electrons can make the transition between the two bands without having to lose or gain momentum in the process. Any necessary loss or gain of momentum must

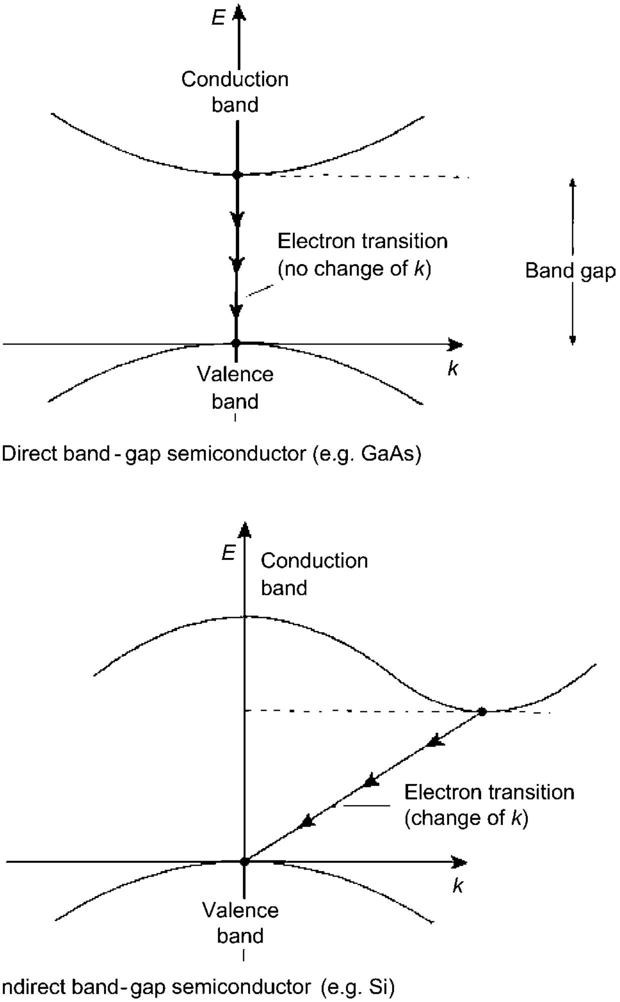

E
Conduction band
Electron transition (no change of k)
Band gap
Valence band
k
(a) Direct band-gap semiconductor (e.g. GaAs)
E
Conduction band
Electron transition (change of k)
Valence band
k
(b) Indirect band-gap semiconductor (e.g. Si)

Figure A2.1.46. Energy level diagrams for direct and indirect band-gap semiconductors.

always involve a third entity, a ‘phonon’ (quantum of vibration) for example, and this renders the transition much less probable. Hence a direct-band-gap material is much more efficient than an indirect-band-gap material (figure A2.1.46(*b*)) and the processes are much faster, leading to higher device bandwidth.

Quite frequently even finer control is required over the value of the band gap and, for this, ‘band-gap engineers’ turn to ternary alloys, i.e. those involving three elements, where the ratio of III to V composition is still approximately 1:1. An example is the range of alloys which is described by the formula $\mathrm{Al}_x\mathrm{Ga}_{1-x}\,\mathrm{As}$. By varying x, one can move along the line between GaAs and AlAs on figure A2.1.47 and thus vary the band gap appropriately. Figure A2.1.47 shows other materials which can be tailored in this way.

One final difficulty is that, in order to grow the required material one needs a substrate on which to grow it, from either the gas phase (gas-phase epitaxy) or the liquid phase (liquid-phase epitaxy), and this requires that the desired material has approximately the same lattice spacing as the substrate. For the example of $\mathrm{Al}_x\mathrm{Ga}_{1-x}\,\mathrm{As}$ there is little difficulty since the GaAs/AlAs line is almost vertical (figure A2.1.47), and thus the lattice spacing always is close to that of GaAs, which can thus be used as the substrate. For other materials, for example, $\mathrm{InAs}_y\mathrm{Sb}_{1-y}$, this is clearly not the case since the InAs/InSb line is almost horizontal. This problem can be solved by going one stage yet further, to quaternary alloys which lie in the regions bounded by the lines in figure A2.1.47. An example of a quaternary alloy is $\mathrm{In}_x\mathrm{Ga}_{1-x}\,\mathrm{As}_y\mathrm{P}_{1-y}$ and one of these alloys is shown marked x on figure A2.1.47. This has a lattice spacing similar to that of InP, which can thus be used as a satisfactory substrate.

Thus a band-gap engineer generally will choose a substrate, then a suitable quaternary alloy which has the required band gap, probably making sure it is a direct-band-gap material, and then grow the semiconductor.

Band-gap engineering is now a sophisticated, and extremely valuable, technology for the provision of materials for optoelectronic devices; their behaviour and performance depend critically on the material from which they are made.

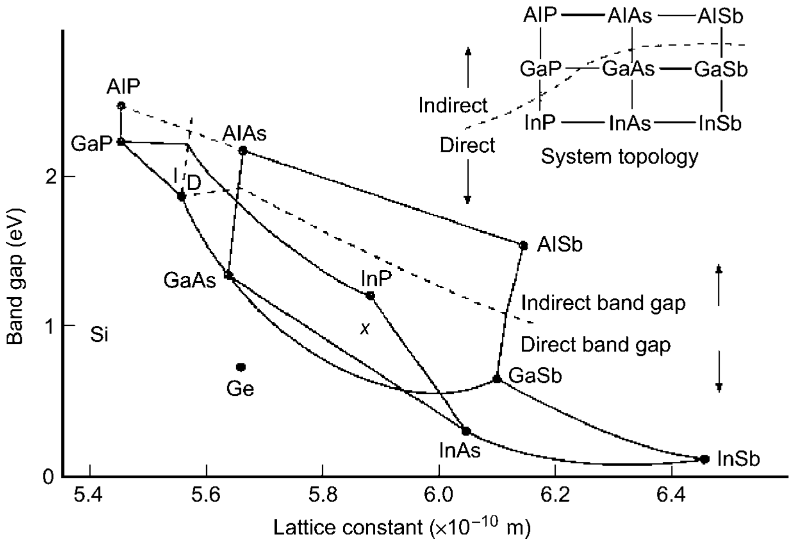


Figure A2.1.47. Alloy band-gap diagram.

A2.1.6.6 Summary

Solids are complex states of matter, with their range of overlapping atomic wavefunctions. Many of the ideas are broadly unfamiliar to a nonspecialist physicist. However, many optoelectronic devices and systems rely on the behaviour of electrons in solids, and we need to have a good understanding of this. In studying optoelectronics, we certainly need to draw repeatedly upon the ideas outlined in this section.

A2.1.7 Nonlinear optics

A2.1.7.1 Introduction

In all of the various discussions concerning the propagation of light in material media so far, we have been dealing with linear processes. By this we mean that a light beam of a certain optical frequency which enters a given medium will leave the medium with the same frequency, although the amplitude and phase of the wave will, in general, be altered.

The fundamental physical reason for this linearity lies in the way in which the wave propagates through a material medium. The effect of the electric field of the optical wave on the medium is to set the electrons of the atoms (of which the medium is composed) into forced oscillation; these oscillating electrons then radiate secondary wavelets (since all accelerating electrons radiate) and the secondary wavelets combine with each other and with the original (primary) wave, to form a resultant wave. Now the important point here is that all the forced electrons oscillate at the same frequency (but differing phase, in general) as the primary, driving wave, and thus we have the sum of the waves all of same frequency, but with different amplitudes and phases.

If two such sinusoids are added together:

$$A_{\mathrm{T}} = a_1 \ \sin(\omega t + \varphi_1) + a_2 \ \sin(\omega t + \varphi_2)$$

and we have, from simple trigonometry:

$$A_{\mathrm{T}} = a_{\mathrm{T}} \ \sin(\omega t + \varphi_{\mathrm{T}})$$

where

$$a_{\mathrm{T}}^2 = a_1^2 + a_2^2 + 2a_1a_2 \ \cos(\varphi_1 - \varphi_2)$$

and

$$\tan\varphi_{\mathrm{T}} = \frac{a_1\sin\varphi_1 + a_2\sin\varphi_2}{a_1\cos\varphi_1 + a_2\cos\varphi_2}.$$

In other words, the resultant is a sinusoid of the same frequency but of different amplitude and phase. It follows, then, that no matter how many more such waves are added, the resultant will always be a wave of the same frequency, i.e.

$$A_{\mathrm{T}} = \sum_{n=0}^{N} \ a_n \sin(\omega t + \varphi_n) = \alpha \sin(\omega t + \beta)$$

where α and β are expressible in terms of a_n and ϕ_n.

It follows, further, that if there are two primary input waves, each will have the effect described above independently of the other, for each of the driving forces will act independently and the two will add to produce a vector resultant. We call this the 'principle of superposition' for linear systems since the resultant effect of the two (or more) actions is just the sum of the effects of each on acting on its own.

This has to be the case whilst the displacements of the electrons from their equilibrium positions in the atoms varies linearly with the force of the optical electric fields. Thus, if we pass into a medium, along the same path, two light waves, of angular frequencies ω_1 and ω_2, emerging from the medium will be two light waves (and only two) with those same frequencies, but with different amplitudes and phases from the input waves.

Suppose now, however, that the displacement of the electrons is *not* linear with the driving force. Suppose, for example, that the displacement is so large that the electron is coming close to the point of breaking free from the atom altogether. We are now in a nonlinear regime. Strange things happen here. For example, a given optical frequency input into the medium may give rise to waves of several different frequencies at the output. Two frequencies ω_1 and ω_2 passing in may lead to, among others, sum and difference frequencies $\omega_1 \pm \omega_2$ coming out.

The fundamental reason for this is that the driving sinusoid has caused the atomic electrons to oscillate nonsinusoidally (figure A2.1.48). Our knowledge of Fourier analysis tells us that any periodic nonsinusoidal function contains, in addition to the fundamental component, components at harmonic frequencies, i.e. integral multiples of the fundamental frequency.

This is a fascinating regime. All kinds of interesting new optical phenomena occur here. As might be expected, some are desirable, some are not. Some are valuable in new applications, some just comprise sources of noise. But to use them to advantage, and to minimize their effects when they are a nuisance, we must understand them better. This we shall now try to do.

A2.1.7.2 Nonlinear optics and optical fibres

Let us begin by summarizing the conditions which give rise to optical nonlinearity.

In a semiclassical description of light propagation in dielectric media, the optical electric field drives the atomic/molecular oscillators of which the material is composed, and these oscillators become secondary radiators of the field; the primary and secondary fields then combine vectorially to form the resultant wave. The phase of this wave (being different from that of its primary) determines a velocity of light different from that of free space, and its amplitude determines a scattering/absorption coefficient for the material.

Nonlinear behaviour occurs when the secondary oscillators are driven beyond the linear response; as a result, the oscillations become nonsinusoidal. Fourier theory dictates that, under these conditions, frequencies other than that of the primary wave will be generated (figure A2.1.48).

The fields necessary to do this depend upon the structure of the material, since it is this which dictates the allowable range of sinusoidal oscillation at given frequencies. Clearly, it is easier to generate large amplitudes of oscillation when the optical frequencies are close to natural resonances, and one expects (and obtains) enhanced nonlinearity there. The electric field required to produce nonlinearity in

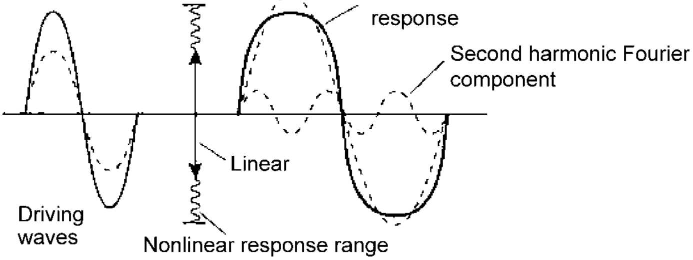


Figure A2.1.48. Nonlinear response to a sinusoidal drive.

material therefore varies widely, from $\sim 10^6$ up to $\sim 10^{11}\,\mathrm{V\,m^{-1}}$, the latter being comparable with the atomic electric field. Even the lower of these figures, however, corresponds to an optical intensity of $\sim 10^9\,\mathrm{W\,m^{-2}}$, which is only achievable practically with laser sources. It is for this reason that the study of nonlinear optics only really began with the invention of the laser, in 1960.

The magnitude of any given nonlinear effect will depend upon the optical intensity, the optical path over which the intensity can be maintained, and the size of the coefficient which characterizes the effect.

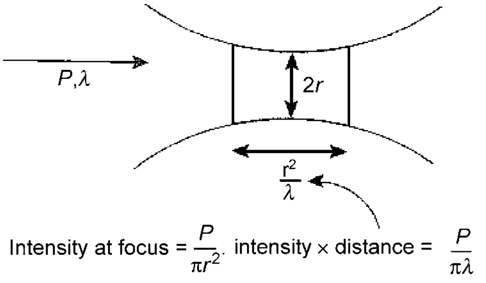
In bulk media, the magnitude of any nonlinearity is limited by diffraction effects. For a beam of power P watts and wavelength λ focused to a spot of radius r, the intensity, $P/\pi r^2$, can be maintained (to within a factor of ~ 2) over a distance $\sim r^2/\lambda$ (Rayleigh distance), beyond which diffraction will rapidly reduce it. Hence the product of intensity and distance is $\sim P/\pi\lambda$, independent of r and of propagation length (figure A2.1.49(*a*)).

However, in an optical fibre the waveguiding properties, in a small diameter core, serve to maintain a high intensity over lengths of up to several kilometers (figure A2.1.49(*b*)). This simple fact allows magnitudes of nonlinearities, in fibres, which are many orders greater than in bulk materials. Further, for maximum overall effect, the various components' effects per elemental propagation distance must add coherently over the total path. This implies a requirement for phase coherence throughout the path which, in turn, implies a single propagation mode: monomode rather than multimode fibres must, in general, be used.

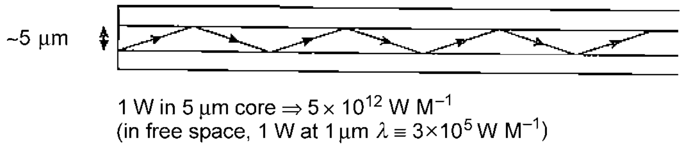
A2.1.7.3 The formalism of nonlinear optics

Under 'normal' propagation conditions we assume a linear relationship between the electric polarization (P) of a medium and the electric field (E) of an optical wave propagating in it, by taking:

$$\chi = \frac{P}{E}$$


P,λ
$2r$
$\frac{r^2}{\lambda}$
Intensity at focus = $\frac{P}{\pi r^2}$, intensity × distance = $\frac{P}{\pi\lambda}$

(*a*) The Rayleigh distance for free-space focusing


~5 μm
1 W in 5 μm core ⇒ 5×10^{12} W M^{-1}
(in free space, 1 W at 1 μm $\lambda \equiv 3\times10^5$ W M^{-1})

(*b*) Nonlinear facility in optical fibres

Figure A2.1.49. The intensity-distance product for nonlinearity.

where χ is the volume susceptibility of the medium, and is assumed constant. The underlying assumption for this is that the separation of atomic positive and negative charges is proportional to the imposed field, leading to a dipole moment per unit volume (P) which is proportional to the field.

However, it is clear that the linearity of this relationship cannot persist for ever-increasing strengths of field. Any resonant physical system must eventually be torn apart by a sufficiently strong perturbing force and, well before such a catastrophe occurs, we expect the separation of oscillating components to vary nonlinearly with the force. In the case of an atomic system under the influence of the electric field of an optical wave, we can allow for this nonlinear behaviour by writing the electric polarization of the medium in the more general form:

$$P(E) = \chi_1 E + \chi_2 E^2 + \chi_3 E^3 + \chi_j E^j + \cdots. \tag{A2.1.34}$$

The value of χ_j (often written $\chi^{(j)}$) decreases rapidly with increasing j for most materials. Also the importance of the j-th term, compared with the first, varies as $(\chi_j/\chi_1)E^{(j-1)}$, and so depends strongly on E. In practice, only the first three terms are of any great importance, and then only for laser-like intensities, with their large electric fields. It is not until one is dealing with power densities of $\sim 10^9\,\mathrm{W\,m^{-2}}$, and fields $\sim 10^6\,\mathrm{V\,m^{-1}}$, that $\chi_2 E^2$ becomes comparable with $\chi_1 E$.

Let us now consider the refractive index of the medium. From elementary electromagnetism we know that:

$$\varepsilon = 1 + \chi, \quad n^2 = \varepsilon$$

where χ is the electric permittivity of the medium and n is its refractive index.

Hence

$$n = (1 + \chi)^{1/2} = \left(1 + \frac{P}{E}\right)^{1/2}$$

i.e.

$$n = (1 + \chi_1 + \chi_2 E + \cdots \chi_j E^{j-1} + \cdots)^{1/2}. \tag{A2.1.35}$$

Hence we note that the refractive index has become dependent on E. The optical wave, in this nonlinear regime, is altering its own propagation conditions as it travels. This is a central feature of nonlinear optics.

A2.1.7.4 Second harmonic generation and phase matching

Probably the most straightforward consequence of nonlinear optical behaviour in a medium is that of the generation of the second harmonic of a fundamental optical frequency. To appreciate this mathematically, let us assume that the electric polarization of an optical medium is quite satisfactorily described by the first two terms of equation (A2.1.34), i.e.:

$$P(E) = \chi_1 E + \chi_2 E^2. \tag{A2.1.36}$$

Before proceeding, there is an important point to make about equation (A2.1.36).

Let us consider the effect of a change in sign of E. The two values of the field, ± E, will correspond to two values of P:

$$P(+E) = \chi_1 E + \chi_2 E^2$$

$$P(-E) = -\chi_1 E + \chi_2 E^2.$$

These two values clearly have different absolute magnitudes. Now if a medium is isotropic (as is the amorphous silica of which optical fibre is made) there can be no directionality in the medium and thus the matter of the sign of E, i.e. whether the electric field points up or down, cannot be of any physical relevance and cannot possibly have any measurable physical effect. In particular, it cannot possibly affect the value of the electric polarization (which is, of course, readily measurable). We should expect that changing the sign of E will merely change the sign of P, but that the magnitude of P will be exactly the same: the electrons will be displaced by the same amount in the opposite direction, all directions being equivalent. Clearly this can only be so if $\chi_2 = 0$. The same argument extended to higher order terms evidently leads us to the conclusion that all even-order terms *must* be zero for amorphous (isotropic) materials, i.e. $\chi_{2m} = 0$. This is a point to remember. The corollary of this argument is that in order to retain any even order terms the medium must exhibit some anisotropy. It must, for example, have a crystalline structure without a centre of symmetry. It follows that equation (A2.1.36) refers to such a medium.

Suppose now that we represent the electric field of an optical wave entering such a crystalline medium by:

$$E = E_0 \ \cos \omega t.$$

Then substituting into equation (A2.1.36) we find

$$P(E) = \chi_1 E_0 \cos \omega t + \frac{1}{2}\chi_2 E_0^2 + \frac{1}{2}\chi_2 E_0^2 \cos 2\omega t.$$

The last term, the second harmonic term at twice the original frequency, is clearly in evidence. Fundamentally, it is due to the fact that it is easier to polarize the medium in one direction than in the opposite direction, as a result of the crystal asymmetry. A kind of ‘rectification’ occurs.

Now the propagation of the wave through the crystal is the result of adding the original wave to the secondary wavelets from the oscillating dipoles which it induces. These oscillating dipoles are represented by P. Thus $\partial^2 P/\partial t^2$ leads to e/m waves, since radiated power is proportional to the acceleration of charges, and waves at all of P’s frequencies will propagate through the crystal.

Suppose now that an attempt is made to generate a second harmonic over a length L of crystal. At each point along the path of the input wave a second harmonic component will be generated. But, since the crystal medium will almost certainly be dispersive, the fundamental and second harmonic components will travel at different velocities. Hence the successive portions of second harmonic component generated by the fundamental will not, in general, be in phase with each other, and thus will not interfere constructively. Hence, the efficiency of the generation will depend upon the velocity difference between the waves.

A rigorous treatment of this process requires a manipulation involving Maxwell’s equations but a semi-analytical treatment which retains a firm grasp of the physics will be given here.

Suppose that the amplitude of the fundamental (driving) wave between distances z and $z + \mathrm{d}z$ along the optical path in the crystal is $e \cos(\omega t - kz)$. Then from equation (A2.1.34), there will be a component of electric polarization (dipole moment per unit volume) of the form: $\chi_2 e^2 \cos^2(\omega t - kz)$ giving a time-varying second harmonic term $(1/2)\chi_2 e^2 \cos 2(\omega t - kz)$, as before. Consider, then, a slab, in the medium, of unit cross-section, and thickness $\mathrm{d}z$ (figure A2.1.50).

For this slab, the dipole moment will be:

$$P = \frac{1}{2}\chi_2 e^2 \ \cos\, 2(\omega t - kz)\mathrm{d}z.$$

Now a time-varying dipole moment represents a movement of charge, and therefore an electric current.

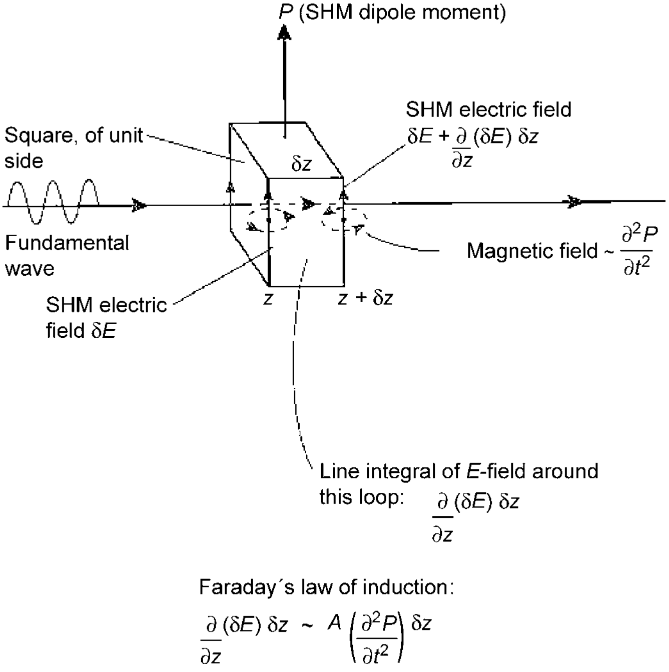

P (SHM dipole moment)
Square, of unit side
δz
SHM electric field δE + ∂/∂z (δE) δz
Fundamental wave
Magnetic field ~ ∂²P/∂t²
SHM electric field δE
z
z + δz
Line integral of E-field around this loop: ∂/∂z (δE) δz
Faraday's law of induction:
∂/∂z (δE) δz ~ A (∂²P/∂t²) δz

Figure A2.1.50. Infinitesimals for second-harmonic generation along the path of the fundamental.

This current will create a magnetic field. A time-varying magnetic field (second derivative $\partial^2 P/\partial t^2$ of the dipole moment) will generate a voltage around any loop through which it threads (Faraday's law of electromagnetic induction). From figure A2.1.50 it can be seen that, if δE is the elemental component of the second harmonic electric field generated by the changing dipole moment in the thin slab, then this voltage is proportional to $\partial(\delta E)/\partial z$. Hence, we have

$$\frac{\partial(\delta E)}{\partial z} = A\frac{\partial^2 P}{\partial t^2} \tag{A2.1.37}$$

where A is a constant.

Hence, from equation (A2.1.37):

$$\frac{\partial(\delta E)}{\partial z} = -A\cdot 2\omega^2\chi_2 e^2 \cos 2(\omega t - kz)\mathrm{d}z.$$

Integrating this with respect to z gives:

$$\delta E = A\frac{\omega^2}{k}\chi_2 e^2 \sin 2(\omega t - kz)\mathrm{d}z$$

and, with $\omega/k = c$, we have:

$$\delta E = Ac\omega\chi_2 e^2 \sin 2(\omega t - kz)dz$$

as the element of the second harmonic electric field generated by the slab between z and $z + \mathrm{d}z$. But the second harmonic component now propagates with wavenumber k_s, say (since the crystal will have a different refractive index at frequency 2ω, compared with that at ω), so when this component emerges from the crystal after a further distance $L - z$, it will have become:

$$\delta E_{\mathrm{L}} = Ac\omega\chi_2 e^2 \sin[2\omega t - 2kz - k_{\mathrm{s}}(L - z)]dz.$$

Hence the total electric field amplitude generated over the length L of crystal will be, on emergence:

$$E_{\mathrm{L}}(2\omega) = \int_0^L Ac\omega\chi_2 e^2 \sin[2\omega t - 2kz - k_{\mathrm{s}}(L - z)]\mathrm{d}z.$$

Performing this integration gives:

$$E_{\mathrm{L}}(2\omega) = Ac\chi_2 e^2 L\omega \frac{\sin\left(k - \frac{1}{2}k_{\mathrm{s}}\right)L}{\left(k - \frac{1}{2}k_{\mathrm{s}}\right)L} \sin\left[2\omega t - (2k + k_{\mathrm{s}})L\right].$$

The intensity of the emerging second harmonic will be proportional to the square of amplitude of this, i.e.

$$I_{\mathrm{L}}(2\omega) = B\chi_2^2 e^4 L^2 \omega^2 \left[\frac{\sin\left(k - \frac{1}{2}k_{\mathrm{s}}\right)L}{\left(k - \frac{1}{2}k_{\mathrm{s}}\right)L}\right]^2$$

where B is another constant. Now the intensity of the fundamental wave is proportional to e^2, so the intensity of the second harmonic is proportional to the square of the intensity of the fundamental, i.e.

$$I_{\mathrm{L}}(2\omega) = B'\chi_2^2 I_{\mathrm{L}}^2(\omega) L^2 \omega^2 \left[\frac{\sin\left(k - \frac{1}{2}k_{\mathrm{s}}\right)L}{\left(k - \frac{1}{2}k_{\mathrm{s}}\right)L}\right]^2 \tag{A2.1.38}$$

where B' is yet another constant. From this we can define an efficiency η_{SHG} for the second harmonic generation process as:

$$\eta_{\mathrm{SHG}} = \frac{I_{\mathrm{L}}(2\omega)}{I_{\mathrm{L}}(\omega)}.$$

Note that η_{SHG} varies as the square of the fundamental frequency and of the length of the crystal; note also that it increases linearly with the power of the fundamental.

From equation (A2.1.38) it is clear that, for maximum intensity, we require that the sinc^2 function has its maximum value, i.e.

$$k_{\mathrm{s}} = 2k_{\mathrm{f}}.$$

This is the *phase matching condition* for second harmonic generation. Now the velocities of the fundamental and the second harmonic are given by:

$$c_{\mathrm{f}} = \frac{\omega}{k_{\mathrm{f}}}, \quad c_{\mathrm{s}} = \frac{2\omega}{k_{\mathrm{s}}}.$$

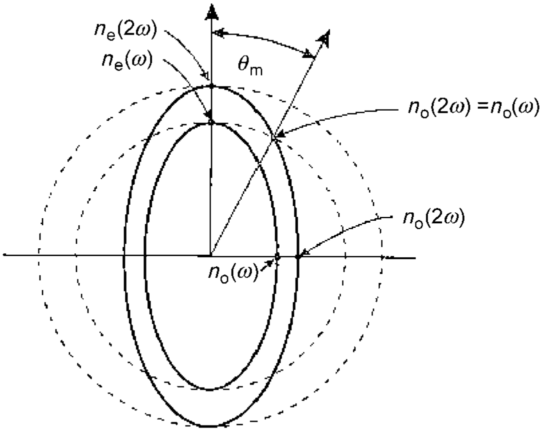
These are equal when $k_{\mathrm{s}} = 2k_{\mathrm{f}}$, so the phase-matching condition is equivalent to a requirement that the two velocities are equal. This is to be expected, since it means that the fundamental generates, at each point in the material, second harmonic components which will interfere constructively. The phase-match condition usually can be satisfied by choosing the optical path to lie in a particular direction within the crystal. It has already been noted that the material must be anisotropic for second harmonic generation to occur; it will also, therefore, exhibit birefringence (section A2.1.3.3). One way of solving the phase-matching problem, therefore, is to arrange that the velocity difference resulting from birefringence is cancelled by that resulting from material dispersion. In a crystal with normal dispersion, the refractive index of both the eigenmodes (i.e. both the ordinary and extraordinary rays) increases with frequency. Suppose we consider the specific example of quartz, which is a positive uniaxial crystal (see section A2.1.3.3). This means that the principal refractive index for the extraordinary ray is greater than that for the ordinary ray, i.e.

$$n_e > n_o.$$

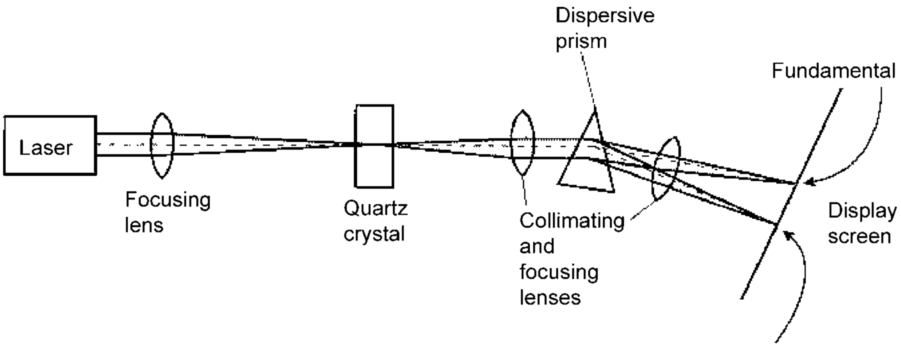
Since quartz is also normally dispersive, it follows that:

$$n_e^{(2\omega)} > n_o^{(\omega)}$$
$$n_o^{(2\omega)} > n_o^{(\omega)}.$$

Hence the index ellipsoids for the two frequencies are as shown in figure A2.1.51(*a*). Now it will be remembered from section A2.1.3.3 that the refractive indices for the 'o' and 'e' rays for any given direction in the crystal are given by the major and minor axes of the ellipse in which the plane normal to


$n_e(2\omega)$
$n_e(\omega)$
θ_m
$n_o(2\omega) = n_o(\omega)$
$n_o(2\omega)$
$n_o(\omega)$

(*a*) Phase matching with the birefringence index ellipsoids


Dispersive prism
Fundamental
Laser
Focusing lens
Quartz crystal
Collimating and focusing lenses
Display screen
Second harmonic

(*b*) Schematic experimental arrangement for SHM generation

Figure A2.1.51. Conditions for second-harmonic generation in quartz.

the direction, and passing through the centre of the index ellipsoid, intersects the surface of the ellipsoid. The geometry (figure A2.1.51(*a*)) thus makes it clear that a direction can be found [10] for which

$$n_{\mathrm{o}}^{(2\omega)}(\vartheta_{\mathrm{m}}) = n_{e}^{(\omega)}(\vartheta_{\mathrm{m}})$$

so SHG phase matching occurs provided that

$$n_{\mathrm{o}}^{(2\omega)} < n_{e}^{(\omega)}.$$

The above is indeed true for quartz over the optical range. Simple trigonometry allows ϑ_{m} to be determined in terms of the principal refractive indices as:

$$\sin^2 \vartheta_{\mathrm{m}} = \frac{(n_{\mathrm{e}}^{(\omega)})^{-2} - (n_{\mathrm{e}}^{(2\omega)})^{-2}}{(n_{\mathrm{o}}^{(2\omega)})^{-2} - (n_{\mathrm{e}}^{(2\omega)})^{-2}}.$$

Hence ϑ_{m} is the angle at which phase matching occurs. It also follows from this that, for second harmonic generation in this case, the wave at the fundamental frequency must be launched at an angle ϑ_{m} with respect to the crystal axis and *must have the 'extraordinary' polarization*; and that the second harmonic component *will appear in the same direction and will have the 'ordinary' polarization*, i.e. the two waves are collinear and have orthogonal linear polarizations! Clearly, other crystal-direction and polarization arrangements also are possible in other crystals.

The required conditions can be satisfied in many crystals but quartz is an especially good one owing to its physical robustness, its ready obtainability with good optical quality and its high optical power-handling capacity.

Provided that the input light propagates along the chosen axis, the conversion efficiency ($\omega \longrightarrow 2\omega$) is a maximum compared with any other path (per unit length) through the crystal. Care must be taken, however, to minimize the divergence of the beam (so that most of the energy travels in the chosen direction) and to ensure that the temperature remains constant (since the birefringence of the crystal will be temperature dependent).

The particle picture of the second harmonic generation process is viewed as an annihilation of two photons at the fundamental frequency, and the creation of one photon at the second harmonic frequency. This pair of processes is necessary in order to conserve energy i.e.

$$2h\nu_{\mathrm{f}} = h(2\nu_{\mathrm{f}}) = h\nu_{\mathrm{s}}.$$

The phase-matching condition is then equivalent to conservation of momentum. The momentum of a photon wave number k is given by:

$$p = \frac{h}{2\pi} k$$

and thus conservation requires that:

$$k_{\mathrm{s}} = 2k_{\mathrm{f}}$$

as in the wave treatment.

Quantum processes which have no need to dispose of excess momentum are again the most probable, and thus this represents the condition for maximum conversion efficiency in the particle picture.

The primary practical importance of second harmonic generation is that it allows laser light to be produced at the higher frequencies, into the blue and ultraviolet, where conditions are not intrinsically favourable for laser action, as was noted earlier (section A2.1.2.4.1). In this context we note again, from equation (A2.1.38), that the efficiency of the generation increases as the square of the fundamental frequency, which is of assistance in producing these higher frequencies.

A2.1.7.5 *Optical mixing*

Optical mixing is a process closely related to second harmonic generation. If, instead of propagating just one laser wave through the same nonlinear crystal, we superimpose two (at different optical frequencies) simultaneously along the same direction, then we shall generate sum and difference frequencies, i.e.

$$E = E_1 \cos \omega_1 t + E_2 \cos \omega_2 t$$

and thus again using equation (A2.1.36):

$$P(E) = \chi_1 (E_1 \cos \omega_1 t + E_2 \cos \omega_2 t) + \chi_2 (E_1 \cos \omega_1 t + E_2 \cos \omega_2 t)^2.$$

This expression for $P(E)$ is seen to contain the term

$$2\chi_2 E_1 E_2 \cos \omega_1 t \cos \omega_2 t = \chi_2 E_1 E_2 \cos(\omega_1 + \omega_2)t + \chi_2 E_1 E_2 \cos(\omega_1 - \omega_2)t$$

giving the required sum and difference frequency terms. Again, for efficient generation of these components, we must ensure that they are phase matched. For example, to generate the sum frequency efficiently we require that:

$$k_1 + k_2 = k_{(1+2)}$$

which is equivalent to

$$\omega_1 n_1 + \omega_2 n_2 = (\omega_1 + \omega_2) n_{(1+2)}$$

where n represents the refractive indices at the suffix frequencies. The condition again is satisfied by choosing an appropriate direction relative to the crystal axes.

This mixing process is particularly useful in the reverse sense. If a suitable crystal is placed in a Fabry–Perot cavity which possesses a resonance at ω_1, say, and is 'pumped' by laser radiation at $\omega_{(1+2)}$, then the latter generates both ω_1 and ω_2. This process is called parametric oscillation: ω_1 is called the signal frequency and ω_2 the idler frequency. It is a useful method for 'down conversion' of an optical frequency i.e. conversion from a higher to a lower value.

The importance of phase matching in nonlinear optics cannot be overstressed. If waves at frequencies different from the fundamental are to be generated efficiently they must be produced with the correct relative phase to allow constructive interference, and this, as we have seen, means that velocities must be equal to allow phase matching to occur. This feature dominates the practical application of nonlinear optics.

A2.1.7.6 *Intensity-dependent refractive index*

It was noted in section A2.1.7.4 that all the even-order terms in expression (A2.1.34) for the nonlinear susceptibility (χ) are zero for an amorphous (i.e. isotropic) medium. This means that in an optical fibre made from amorphous silica, we can expect that $\chi_{(2m)} = 0$, so it will not be possible to generate a second harmonic according to the principles outlined in section A2.1.7.4. (However, second harmonic generation has been observed in fibres [11] for reasons which took some time to understand!) It is possible to generate a third harmonic, however, since to a good approximation the electric polarization in the fibre can be expressed by:

$$P(E) = \chi_1 E + \chi_3 E^3. \qquad \text{(A2.1.39)}$$

Clearly, though, if we wish to generate the third harmonic efficiently we must again phase match it with the fundamental, and this means that somehow we must arrange for the two relevant velocities to be equal, i.e. $c_\omega = c_{3\omega}$. This is very difficult to achieve in practice, although it has been done.

There is, however, a more important application of equation (A2.1.39) in amorphous media. It is clear that the effective refractive index in this case can be written:

$$n_e = (1 + \chi_1 + \chi_3 E^2)^{1/2}$$

and, if $\chi_1, \chi_3 E^2 \ll 1$,

$$n_e \approx 1 + \frac{1}{2}\chi_1 + \frac{1}{2}\chi_3 E^2.$$

Hence:

$$n_e = n_o + \frac{1}{2}\chi_3 E^2 \qquad \text{(A2.1.40}a\text{)}$$

where n_o is the 'normal', linear refractive index of the medium. But we know that the intensity (power/unit area) of the light is proportional to E^2, so that we can write:

$$n_e = n_o + n_2 I \qquad \text{(A2.1.40}b\text{)}$$

where n_2 is a constant for the medium. Equation (A2.1.40*b*) is very important and has a number of practical consequences. We can see immediately that it means that the refractive index of the medium depends upon the intensity of the propagating light: the light is influencing its own velocity as it travels.

In order to fix ideas to some extent let us consider some numbers for silica. For amorphous silica $n_2 \sim 3.2 \times 10^{-20}\,\mathrm{m^2\,W^{-1}}$, which means that a 1% change in refractive index (readily observable) will occur for an intensity $\sim 5 \times 10^{17}\,\mathrm{W\,m^{-2}}$. For a fibre with a core diameter $\sim 5\,\mu\mathrm{m}$ this requires an optical power level of 10 MW. Peak power levels of this magnitude are readily obtainable, for short durations, with modern lasers.

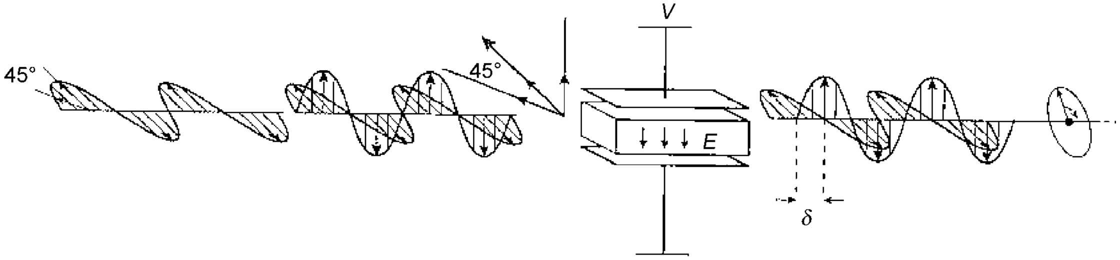
It is interesting to note that this phenomenon is another aspect of the electro-optic effect. Clearly the refractive index of the medium is being altered by an electric field. This will now be considered in more detail.

A2.1.7.7 The electro-optic effect

When an electric field is applied to an optical medium the electrons suffer restricted motion in the direction of the field, when compared with that orthogonal to it. Thus the material becomes linearly birefringent in response to the field. This is known as the electro-optic effect.

Consider the arrangement of figure A2.1.52. Here we have incident light which is linearly polarized at 45° to an electric field and the field acts on a medium transversely to the propagation direction of the light. The field-induced linear birefringence will cause a phase displacement between components of the incident light which lie, respectively, parallel and orthogonal to the field; hence the light will emerge elliptically polarized.

A (perfect) polarizer placed with its acceptance direction parallel with the input polarization direction will of course, pass all the light in the absence of a field. When the field is applied, the fraction of light power which is passed will depend upon the form of the ellipse, which in turn depends upon the phase delay introduced by the field. Consequently, the field can be used to modulate the intensity of the light, and the electro-optic effect is, indeed, very useful for the modulation of light. The phase delay introduced may be proportional either to the field (Pockels effect) or to the square of the field (Kerr effect). All materials manifest a transverse Kerr effect. Only crystalline materials can manifest any



Linear polarization becomes elliptical by passing through an electro-optic medium with applied field E

Figure A2.1.52. The electro-optic effect.

kind of Pockels effect, or longitudinal (***E*** field parallel with propagation direction) Kerr effect. The reason for this is physically quite clear. If a material is to respond linearly to an electric field, the effect of the field must change sign when the field changes sign. This means that the medium must be able to distinguish (for example) between 'up' (positive field) and 'down' (negative field). But it can only do this if it possesses some kind of directionality in itself, otherwise all field directions must be equivalent in their physical effects. Hence, in order to make the necessary distinction between up and down, the material must possess an intrinsic asymmetry, and hence must be crystalline. By a similar argument a longitudinal E field can only produce a directional effect orthogonally to itself (i.e. in the direction of the optical electric field) if the medium is anisotropic (i.e. crystalline) for otherwise all transverse directions will be equivalent. In addition to the modulation of light (phase or intensity/power) it is clear that the electro-optic effect could be used to measure an electric field and/or the voltage which gives rise to it.

A2.1.7.8 Optical Kerr effect

The normal electro-optic Kerr effect is an effect whereby an electric field imposed on a medium induces a linear birefringence with slow axis parallel with the field (figure A2.1.53(*a*)). The value of the induced birefringence is proportional to the square of the electric field. In the optical Kerr effect the electric field involved is that of an optical wave, and thus the birefringence probed by one wave may be that produced by another (figure A2.1.53(*b*)).

The phase difference introduced by an electric field E over an optical path L is given by:

$$\Delta\varphi = \frac{2\pi}{\lambda}\Delta n L$$

where $\Delta n = KE^2$, K being the Kerr constant.

Now from equations (A2.1.40*a*) and (A2.1.40*b*) we have:

$$\Delta n = n_2 I = \frac{1}{2}\chi_3 E^2 = KE^2. \tag{A2.1.41}$$

From elementary electromagnetism we know that:

$$I = c\varepsilon E^2.$$

Hence we have, from equation (A2.1.41)

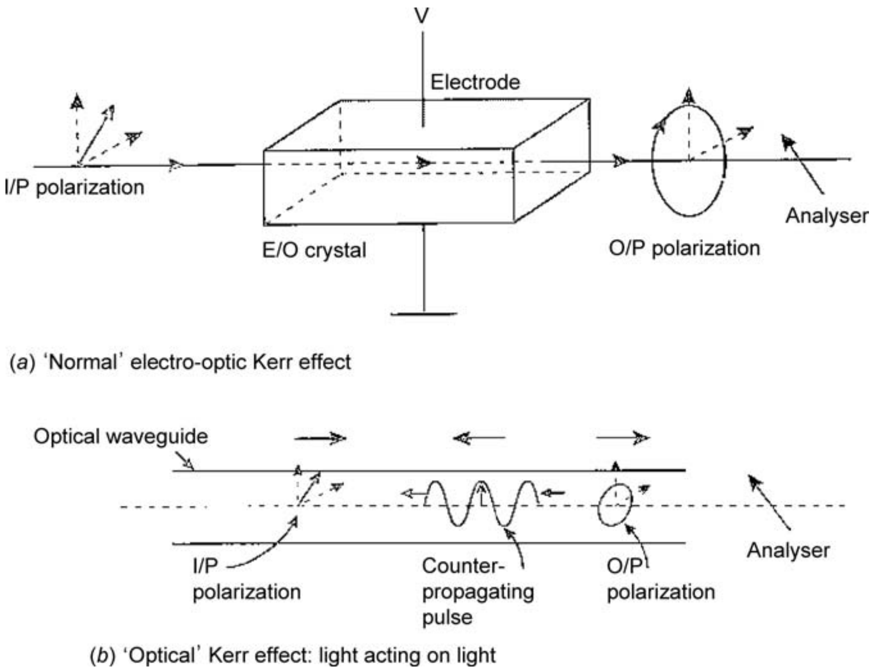

V
Electrode
I/P polarization
E/O crystal
O/P polarization
Analyser
(a) 'Normal' electro-optic Kerr effect
Optical waveguide
I/P
polarization
Counter-
propagating
pulse
O/P
polarization
Analyser
(b) 'Optical' Kerr effect: light acting on light

Figure A2.1.53. 'Normal' and 'optical' Kerr effects.

$$K = n_2 c\varepsilon = \frac{1}{2}\chi_3$$

showing that the electro-optic effect, whether the result of an optical or an external electric field, is a nonlinear phenomenon, depending on χ_3. Using similar arguments it can easily be shown that the electro-optic Pockels effect also is a nonlinear effect, depending on χ_2.

The optical Kerr effect has several other interesting consequences. One of these is self-phase modulation (SPM), which is the next topic for consideration.

A2.1.7.9 Self-phase modulation

The fact that refractive index can be dependent on optical intensity clearly has implications for the phase of the wave propagating in nonlinear medium. We have:

$$\varphi = \frac{2\pi}{\lambda} nL.$$

Hence for $n = n_o + n_2 I$,

$$\varphi = \frac{2\pi L}{\lambda}(n_o + n_2 I).$$

Suppose now that the intensity is a time-dependent function $I(t)$. It follows that φ also will be time dependent, and, since:

$$\omega = \frac{\mathrm{d}\varphi}{\mathrm{d}t}$$

the frequency spectrum will be changed by this effect, which is known as self-phase modulation (SPM).

In a dispersive medium a change in the spectrum of a temporally varying function (e.g. a pulse) will change the shape of the function. For example, pulse broadening or *pulse compression* can be obtained under appropriate circumstances. To see this, consider a Gaussian pulse (figure A2.1.54(*a*)). The Gaussian shape modulates an optical carrier of frequency ω_0, say, and the new instantaneous frequency becomes

$$\omega' = \omega_o + \frac{\mathrm{d}\varphi}{\mathrm{d}t}$$

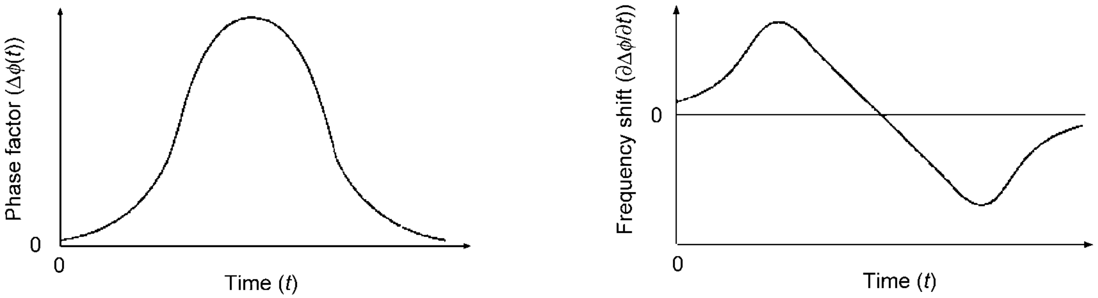
If the pulse is propagating in the Oz direction:

$$\varphi = -\frac{2\pi z}{\lambda}(n_o + n_2 I) \qquad \text{(A2.1.42}a\text{)}$$

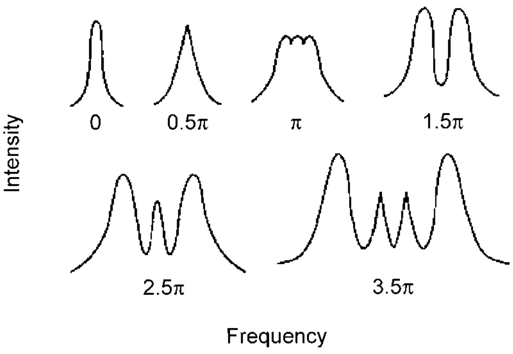
and we have

$$\omega' = \omega_o - \frac{2\pi z}{\lambda} n_2 \frac{\mathrm{d}I}{\mathrm{d}t}. \qquad \text{(A2.1.42}b\text{)}$$

At the leading edge of the pulse $\mathrm{d}I/\mathrm{d}t > 0$, hence



(*a*) Intensity-dependent phase factor for a Gaussian pulse (*b*) The instantaneous frequency shift for (*a*)



(*c*) Frequency spectra for (*a*) designated by maximum phase shift, at peak

Figure A2.1.54. Self-phase modulation for a Gaussian pulse.

$$\omega' = \omega - \omega_{\mathrm{I}}(t).$$

At the trailing edge

$$\frac{\mathrm{d}I}{\mathrm{d}t} < 0$$

and

$$\omega' = \omega + \omega_{\mathrm{I}}(t).$$

Hence the pulse is now ‘chirped’, i.e. the frequency varies across the pulse. Figure A2.1.54(*b*) shows an example of this effect.

Suppose, for example, a pulse from a mode-locked Argon laser, initial width 180 ps, is passed down 100 m of optical fibre. As a result of SPM the frequency spectrum is changed by the propagation. Figure A2.1.54(*c*) shows how the spectrum varies as the initial peak power of the pulse is varied. The peak power will lead to a peak phase change, according to equation (A2.1.42*a*) and this phase change is shown for each of the spectra. It can be seen that the initial spectrum ($\Delta\varphi = 0$) is just due to the modulation of the optical sinusoid (Fourier spectrum of a Gaussian pulse) and, as the value of $\Delta\varphi$ increases, the first effect is a broadening. At $\Delta\varphi = 1.5\pi$ the spectrum has split into two clear peaks, corresponding to the frequency shifts at the back and front edges of the pulse. The spectra then develop multiple peaks.

It is important to realize that this does not necessarily change the shape of the pulse envelope, just the optical frequency within it. However, if the medium through which the pulse is passing is dispersive, the pulse shape will change. This is an interesting possibility and it leads to the phenomenon of soliton propagation [12].

A2.1.7.10 Summary

We have seen in this chapter that nonlinear optics has its advantages and disadvantages. When it is properly under control it can be enormously useful; but on other occasions it can intrude, disturb and degrade.

The processes by which light waves produce light waves of other frequencies need very high optical electric fields and thus high peak intensities. It was for this reason that nonlinear optics only became a serious subject with the advent of the laser. Optical fibres provide a convenient means by which peak intensities can be maintained over relatively long distances, and are thus very useful media for the study and control of nonlinear optical effects.

We must also remember that in order to cause one optical frequency component to generate another, the second must be generated in phase with itself along the generation path: phase matching is an important feature of such processes.

We have seen how the effects which occur when electrons are stretched beyond the comfortable sinusoidal oscillations in their atoms or molecules can yield useful extra optical waves (SHM) and can influence their own propagation conditions (optical Kerr effect, SPM).

Finally we can also use light to alter, permanently or semi-permanently, the optical properties of a medium, and thus provide the means whereby a new class of optical components, especially fibre components, can be fabricated.

There is a wealth of potential here. The exploration of possibilities for nonlinear optics, especially in regard to new, natural or synthetic optical materials (e.g. organics, high T_{c} superconductors) etc, has not even really begun. The prospects, for example, for new storage media, fast switching of light by light, and three-dimensional television, which will be opened up in the future by such materials, are intriguing, and it could well be that nonlinear optical technology soon will become a powerful subject in its own right.

A2.1.8 Conclusions

In this chapter, the aim has been to emphasize the principles which comprise the physical basis of optoelectronics. A thorough grasp of these principles should facilitate a better appreciation of the more advanced treatments of components and systems in later chapters of this Handbook.

Acknowledgments

The majority of the material presented in this chapter first appeared in 'Essentials of Optoelectronics' (Rogers), published by Chapman and Hall, 1997, and is included here with permission.

References

[1] Zemansky M W 1968 *Heat and Thermodynamics* 5th edn (New York: McGraw Hill) chapter 6
[2] Nye J F 1976 *Physical Properties of Crystals* (Oxford: Clarendon) chapter 2
[3] Jones, R C 1941–1956 'A new calculus for the treatment of optical systems', *J. Opt. Soc. Am.* **31** **46** 234–241
[4a] Born M and Wolf E 1975 *Principles of Optics* 5th edn (Oxford: Pergamon) p 30
[4b] Born M and Wolf E 1975 *Principles of Optics* 5th edn (Oxford: Pergamon) p 40
[5] Jerrard H G 1954 Transmission of light through optically active media *J. Opt. Soc. Am.* **44** 634–664
[6] Syms R and Cozens J 1992 *Optical Guided Waves and Devices* (New York: McGraw-Hill) chapter 6
[7] Kaplan W 1981 *Advanced Mathematics for Engineers* (Reading, MA: Addison-Wesley) chapter 12
[8] Adams M J 1981 *An Introduction to Optical Waveguides* (New York: Wiley) chapter 7
[9] Kittel C 1968 *Introduction to Solid State Physics* 3rd edn.(New York: Wiley)
[10] Nye J F 1976 *Physical Properties of Crystals* (Oxford: Clarendon) chapter 13
[11] Fujii Y, Kawasaki B S, Hill K O and Johnson D C 1980 Sum-frequency light generation in optical fibres *Opt. Lett.* **5** 48
[12] Mollenauer L F, Stolen R H and Gordon J P 1980 Experimental observation of picosecond narrowing and solitons in optical fibres *Phys. Rev. Lett.* **45** 1095

Further reading

Siegman A E 1986 *Lasers* (Mills Valley, CA: University Science Books)

Solymar L and Walsh D 1993 *Lectures on the Electrical Properties of Materials* 5th edn (Oxford Science Publications)

Sze S M 1981 *Physics of Semiconductor Devices* 2nd edn (New York: Wiley)

Collett E 1993 *Polarised Light: Fundamentals and Applications*, (New York: Dekker)

Kliger D S Lewis, J W and Randall, C E 1990 *Polarised Light in Optics and Spectroscopy* (New York: Academic)

Shurchiff W A 1962 *Polarised Light: Production and Use* (Cambridge, MA: Harvard University Press. (an excellent introduction)

Hecht E 1987 *Optics* (Reading, MA: Addison-Wesley) 2nd edn chapter 12

Marz R 1995 *Integrated Optics: Design and Modelling*, (Boston, MA: Artech House)

Najafi S I 1992 *Introduction to Glass Integrated Optics* (Boston MA: Artech House)

Midwinter J E 1978 *Optical Fibres for Transmission* (New York: Wiley)

Agrawal G P 1989 *Nonlinear Fiber Optics* (New York: Academic)

Boyd R W 1992 *Nonlinear Optics* (New York: Academic)

Guenther R D 1990 *Modern Optics* (New York: Wiley) chapter 15

Andonovic I and Uttamchandani D 1989 Optical information processing, storage media integrated optics polarimeters *Principles of Modern Optical Systems* (Boston, MA: Artech House)

Bjarklev A 1993 *Optical Fiber Amplifiers* (Boston, MA: Artech House)

Blaker J W and Rosenblum W B 1993 *Optics: An Introduction for Students of Engineering* (New York: Macmillan) (Instruments interferometry and holography)

Dakin J P 1990 *The Distributed Fibre-optic Sensing Handbook* Springer Berlin

Grattan K T V and Meggitt, B T 1995 Interferometric sensors, optical-fibre current measurement, distributed optical-fibre sensors *Optical-Fibre Sensor Technology* (London: Chapman and Hall)

Keiser G 1991 *Optical Fiber Communications* (New York: McGraw-Hill)

Lefevre H 1993 *The Fiber-Optic Gyroscope* (Boston MA: Artech House)

Ryan S 1995 *Coherent Lightwave Communications Systems* (Boston MA: Artech House)

Heavens O S and Ditchburn R W 1991 *Insight into Optics* (New York: Wiley)

A2.2

Basic concepts in photometry, radiometry and colorimetry

Yoshi Ohno

A2.2.1 Introduction

The term *photometry* refers to measurement of quantities for optical radiation as evaluated according to a standardized human eye response, and therefore, is limited to the visible spectral region (360–830 nm) [1]. Photometry uses either optical radiation detectors constructed to mimic the spectral response of the eye or spectroradiometry coupled with appropriate calculations for weighting by the spectral response of the eye. Typical photometric units include the lumen (luminous flux), the candela (luminous intensity), the lux (illuminance) and the candela per square metre (luminance). On the other hand, measurement of optical radiation at all wavelengths (approximately in the range from 10 nm to 1000 μm including ultraviolet, visible and infrared) is referred to as *radiometry*. The official definition of radiometry [1] is measurement of the quantities associated with radiant energy. Typical radiometric units include the watt (radiant flux), watt per steradian (radiant intensity), watt per square metre (irradiance) and watt per square metre per steradian (radiance). Radiometry often involves spectrally resolved measurements of these quantities as well as spectrally integrated measurements. Similar to photometry, measurement of colour of light sources and objects also deals with broadband measurement of the visible radiation and is referred to as *colorimetry*. Colorimetry is ascribed to measurement of light spectra weighted by three standardized spectral weighting functions, one of which is identical to the standardized human eye response used in photometry.

Photometry and colorimetry are essential for evaluation of light sources used for lighting, signalling, displays and other applications where light is seen by the human eye. Light-emitting diodes, for example, are now produced in all colour ranges and expected to gain wide acceptance in many applications. This chapter focuses on fundamentals of photometry, but introduces radiometry and colorimetry also because radiometry is closely related to photometry, and colorimetry is also important for evaluation of optoelectronic light sources. Some of the presented materials in this chapter are from [2] by the same author. The terminology used in this chapter follows international standards and recommendations [1,3,4].

A2.2.2 Basis of physical photometry

A2.2.2.1 Visual response

The primary aim of photometry is to measure light (visible optical radiation) in such a way that the results correlate with what the visual sensation is to a normal human observer exposed to that radiation.

Until about 1940, visual comparison techniques of measurements were predominant in photometry. The intensity of one light source is matched to the intensity of another light source using human eyes. In modern photometric practice, measurements are made with photodetectors. This is referred to as physical photometry. In order to achieve the aim of photometry, one must take into account the characteristics of human vision. The relative spectral responsivity of the human eye was first defined by the Commission Internationale de l'Éclairage (CIE) in 1924 [5]. It is called *the spectral luminous efficiency for photopic vision*, with a symbol $V(\lambda)$, defined in the domain from 360 to 830 nm, and is normalized to one at its peak, 555 nm (figure A2.2.1). This model has gained wide acceptance. The values were republished by CIE in 1983 [6], and adopted by Comité International des Poids et Mesures (CIPM) in 1983 [7] to supplement the 1979 definition of the candela. The tabulated values of the function at 1 nm increments are available in [6–8]. In most cases, the region from 380 to 780 nm suffices for calculation with negligible errors because the value of the $V(\lambda)$ function falls below 10^{-4} outside this region.

As specified in the definition of the candela by Conférence Générale des Poids et Mesures (CGPM) in 1979 [9] and a supplemental document from CIPM in 1982 [10], a photometric quantity X_v is now defined in relation to the corresponding radiometric quantity $X_{e,\lambda}$ by the equation:

$$X_v = K_m \int_{360\,\mathrm{nm}}^{830\,\mathrm{nm}} X_{e,\lambda} V(\lambda)\,\mathrm{d}\lambda. \tag{A2.2.1}$$

The constant, K_m, relates the photometric quantities and radiometric quantities, and is called the *maximum spectral luminous efficacy (of radiation) for photopic vision*. The value of K_m is given in the 1979 definition of candela, which defines the spectral luminous efficacy of radiation at the frequency 540×10^{12} Hz (at the wavelength 555.016 nm in standard air) to be $683\,\mathrm{lm\,W^{-1}}$. Note that this is not exactly at the peak of the $V(\lambda)$, 555 nm. The value of K_m is calculated as $683 \times V(555.000\,\mathrm{nm})/V(555.016\,\mathrm{nm}) = 683.002\,\mathrm{lm\,W^{-1}}$ [6]. K_m is normally rounded to $683\,\mathrm{lm\,W^{-1}}$ with negligible errors.

It should be noted that the $V(\lambda)$ function is defined for the *CIE standard photometric observer for photopic vision*, which assumes additivity of sensation and a 2° field of view at relatively high luminance levels (higher than approximately $1\,\mathrm{cd\,m^{-2}}$). The human vision in this level is called *photopic* vision. The spectral responsivity of human eyes deviates significantly at very low levels of luminance (at luminance levels less than approximately $10^{-3}\,\mathrm{cd\,m^{-2}}$) when the rods in the eyes are the dominant receptors. This type of vision is called *scotopic* vision. Its spectral responsivity, peaking at 507 nm, is designated as $V'(\lambda)$, and was defined by CIE in 1951 [11], recognized by CIPM in 1976 [12] and republished by CIPM in 1983 [7]. The human vision in the region between photopic vision and scotopic vision is called

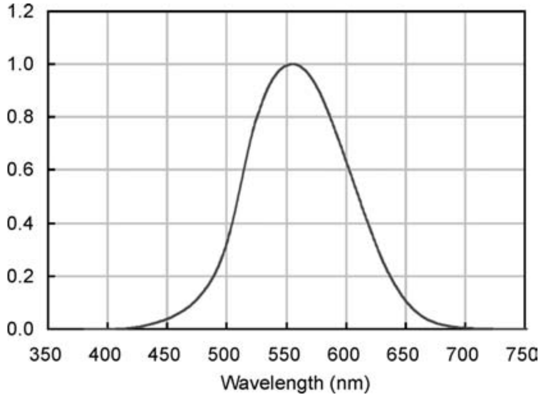


Figure A2.2.1. CIE $V(\lambda)$ function.

mesopic vision. While there have been active researches in this area [13], there is no internationally accepted spectral luminous efficiency function for the mesopic region yet. In current practice, almost all photometric quantities are given in terms of photopic vision, even at such low light levels. Quantities in scotopic vision are seldom used except for special calculations for research purposes. Further details of the contents in this section are available in [6].

A2.2.2.2 Photometric base unit, the candela

The history of photometric standards dates back to the early nineteenth century, when the intensity of light sources was measured in comparison with a standard candle using visual bar photometers [14]. At that time, the flame of a candle was used as a unit of luminous intensity that was called *the candle*. The old name for luminous intensity 'candle power' came from this origin. Standard candles were gradually superseded by flame standards of oil lamps, and in 1920, the unit of luminous intensity, recognized as *the international candle*, was adopted by the CIE. In 1948, it was adopted by the CGPM with a new Latin name 'candela' defined as the luminous intensity of a platinum blackbody at its freezing temperature under specified geometry. Although the 1948 definition served to establish the uniformity of photometric measurements in the world, difficulties in fabricating the blackbodies and in improving accuracy were addressed. In 1979, the candela was redefined in relation to the optical power, watt, so that complicated source standards would not be necessary. The current definition of the candela adopted in 1979 by the CGPM [9] is

> "The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of (1/683) watt per steradian."

The value of K_m ($683\,\mathrm{lm\,W^{-1}}$) was determined in such a way that the consistency from the prior unit was maintained, and was determined based on the measurements by several national laboratories. Technical details on this redefinition of the candela are reported in [15, 16]. This 1979 redefinition of the candela has enabled the derivation of the photometric units from the radiometric units using various techniques.

A2.2.3 Quantities and units in photometry and radiometry

In 1960, the Système International (SI) was established, and the candela became one of the seven SI base units [17]. For further details on the SI, references [17–20] can be consulted. Several quantities and units, defined in different geometries, are used in photometry and radiometry. Table A2.2.1 lists the photometric quantities and units, along with corresponding radiometric quantities and units.

While the candela is the SI base unit, the luminous flux (lumen) is perhaps the most fundamental photometric quantity, as the other photometric quantities are defined in terms of lumen with an appropriate geometric unit. The definitions of these photometric quantities are described later. The descriptions given here are occasionally simplified from the definitions given in official reference [1] for easier understanding. Refer to this reference for official, rigorous definitions.

A2.2.3.1 Radiant flux and luminous flux

Radiant flux (also called *optical power* or *radiant power*) is the energy Q (in joules) radiated by a source per unit of time, expressed as

$$\Phi = \frac{\mathrm{d}Q}{\mathrm{d}t}. \tag{A2.2.2}$$

The unit of radiant flux is the watt ($\mathrm{W} = \mathrm{J\,s^{-1}}$).

Table A2.2.1. Quantities and units used in photometry and radiometry.

Photometric quantity	Unit	Relationship with lumen	Radiometric quantity	Unit
Luminous flux	lm (lumen)		Radiant flux	W (watt)
Luminous intensity	cd (candela)	$\mathrm{lm\,sr^{-1}}$	Radiant intensity	$\mathrm{W\,sr^{-1}}$
Illuminance	lx (lux)	$\mathrm{lm\,m^{-2}}$	Irradiance	$\mathrm{W\,m^{-2}}$
Luminance	$\mathrm{cd\ m^{-2}}$	$\mathrm{lm\,sr^{-1}\,m^{-2}}$	Radiance	$\mathrm{W\,sr^{-1}\,m^{-2}}$
Luminous exitance	$\mathrm{lm\ m^{-2}}$		Radiant exitance	$\mathrm{W\,m^{-2}}$
Luminous exposure	lx s		Radiant exposure	$\mathrm{W\,m^{-2}\,s}$
Luminous energy	lm s		Radiant energy	J (joule)
Total luminous flux	lm (lumen)		Total radiant flux	W (watt)
Colour temperature	K (kelvin)		Radiance temperature	K (kelvin)

Luminous flux (Φ_v) is the time rate of flow of light as weighted by $V(\lambda)$. The unit of luminous flux is the lumen (lm). It is defined as

$$\Phi_\mathrm{v} = K_\mathrm{m} \int_\lambda \Phi_{\mathrm{e},\lambda} V(\lambda)\,\mathrm{d}\lambda \tag{A2.2.3}$$

where $\Phi_{\mathrm{e},\lambda}$ is the spectral concentration of radiant flux as a function of wavelength λ. The term, luminous flux, is often used in the meaning of total luminous flux (see A2.2.3.8) in photometry.

A2.2.3.2 Radiant intensity and luminous intensity

Radiant intensity (I_e) or *luminous intensity* (I_v) is the radiant flux (luminous flux) from a point source emitted per unit solid angle in a given direction, as defined by

$$I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} \tag{A2.2.4}$$

where $\mathrm{d}\Phi$ is the radiant flux or luminous flux leaving the source and propagating in an element of solid angle $\mathrm{d}\Omega$ containing the given direction. The unit of radiant intensity is $\mathrm{W\,sr^{-1}}$, and that of luminous intensity is the candela ($\mathrm{cd} = \mathrm{lm\,sr^{-1}}$) (figure A2.2.2).

The radiant intensity or luminous intensity of a real light source varies with the direction of emission, and is specified or measured for given direction(s) from the light source. In real measurements, the solid angle $\mathrm{d}\Omega$ will be a finite solid angle defined by the area A of the detector surface and distance r from the source (see figure A2.2.3), with $\mathrm{d}\Phi$ being the flux incident on the detector surface. The distance r should be large enough so that the source can be assumed as a point source (far field condition) and the detector area A should be small enough so that $\mathrm{d}\Omega$ is considered negligible. Measurement results tend to vary depending on geometrical conditions if these conditions are not met. See also section A2.2.4.1

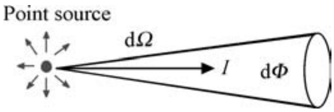


Figure A2.2.2. Radiant intensity and luminous intensity.

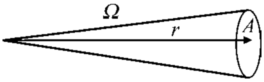


Figure A2.2.3. Solid angle.

for practical aspects of the geometry and reference [30] for the specific measurement geometries recommended for measurement of luminous intensity of LEDs.

Solid angle

The solid angle (Ω) of a cone is defined as the ratio of the area (A) cut out on a spherical surface (with its centre at the apex of that cone) to the square of the radius (r) of the sphere, as given by

$$\Omega = \frac{A}{r^2}. \tag{A2.2.5}$$

The unit of solid angle is steradian (sr), which is a dimensionless unit.

A2.2.3.3 Irradiance and illuminance

Irradiance (E_e) or *illuminance* (E_v) is the density of incident radiant flux or luminous flux at a point on a surface, and is defined as radiant flux or luminous flux per unit area, as given by

$$E = \frac{\mathrm{d}\Phi}{\mathrm{d}A} \tag{A2.2.6}$$

where $\mathrm{d}\Phi$ is the radiant flux or luminous flux incident on an element $\mathrm{d}A$ of the surface containing the point. The unit of irradiance is $\mathrm{W\,m^{-2}}$, and that of illuminance is lux ($\mathrm{lx = lm\,m^{-2}}$) (figure A2.2.4).

By definition, $\mathrm{d}\Phi$ incident on $\mathrm{d}A$ from any angles within the 2π solid angle above the surface contributes to E. Thus, illuminance meters are normally designed to receive light from 2π solid angle (± 90°) with its angular responsivity tailored to follow the cosine function response.

A2.2.3.4 Radiance and luminance

Radiance (L_e) or *luminance* (L_v) is the radiant flux or luminous flux per unit solid angle emitted from a surface element in a given direction, per unit projected area of the surface element perpendicular to the direction (figure A2.2.5). The unit of radiance is $\mathrm{W\,sr^{-1}\,m^{-2}}$, and that of luminance is $\mathrm{cd\,m^{-2}}$. These quantities are defined by

$$L = \frac{\mathrm{d}^2\Phi}{\mathrm{d}\Omega\,\mathrm{d}A\cos\theta} \tag{A2.2.7}$$

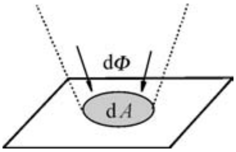


Figure A2.2.4. Irradiance and illuminance.

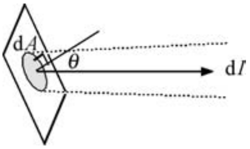


Figure A2.2.5. Radiance and luminance.

where dΦ is the radiant flux (luminous flux) emitted (reflected or transmitted) from the surface element and propagating in the solid angle dΩ containing the given direction. dA is the area of the surface element and θ the angle between the normal to the surface element and the direction of the beam. The term d$A \cos\theta$ gives the projected area of the surface element perpendicular to the direction of measurement.

In photometry, luminance is an important parameter in a sense that it represents how bright objects look to the human eyes. Luminance is important, e.g. to specify the brightness of visual displays. Luminance meters are normally constructed using an imaging optics (like a camera lens) to focus on an object surface and accept light only from a given cone angle (e.g. 0.3°, 1°, 3°, etc) from the meter.

A2.2.3.5 Radiant exitance and luminous exitance

Radiant exitance (M_e) or *luminous exitance* (M_v) is defined to be the density of radiant flux or luminous flux leaving a surface at a point. The unit of radiant exitance is $\mathrm{W\,m^{-2}}$ and that of luminous exitance is $\mathrm{lm\,m^{-2}}$ (but it is not lux) (figure A2.2.6). These quantities are defined by

$$M = \frac{\mathrm{d}\Phi}{\mathrm{d}A} \tag{A2.2.8}$$

where dΦ is the radiant flux or luminous flux leaving the surface element. Luminous exitance is rarely used in the general practice of photometry.

A2.2.3.6 Radiant exposure and luminous exposure

Radiant exposure (H_e) or *luminous exposure* (H_v) is the time integral of irradiance $E_e(t)$ or illuminance $E_v(t)$ over a given duration Δt, as defined by

$$H = \int_{\Delta t} E(t)\,\mathrm{d}t. \tag{A2.2.9}$$

The unit of radiant exposure is $\mathrm{J\,m^{-2}}$, and that of luminous exposure is lux second (lx s). These quantities are often used for pulsed radiation.

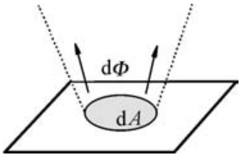


Figure A2.2.6. Radiant exitance and luminous exitance.

A2.2.3.7 Radiant energy and luminous energy

Radiant energy (Q_e) or *luminous energy* (Q_v) is the time integral of the radiant flux $\Phi_e(t)$ or luminous flux $\Phi_v(t)$ over a given duration Δt, as defined by

$$Q = \int_{\Delta t} \Phi(t)\,\mathrm{d}t. \tag{A2.2.10}$$

The unit of radiant energy is joule (J), and that of luminous energy is lumen second (lm s). These quantities are often used for pulsed radiation. Luminous energy is also called *quantity of light*, which is listed in [1].

A2.2.3.8 Total radiant flux and total luminous flux

Total radiant flux or *total luminous flux* (Φ_v) is the geometrically total radiant flux or luminous flux of a light source. It is defined as

$$\Phi = \int_{\Omega} I\,\mathrm{d}\Omega \tag{A2.2.11}$$

or

$$\Phi = \int_{A} E\,\mathrm{d}A \tag{A2.2.12}$$

where I is the radiant or luminous intensity distribution of the light source and E the irradiance or illuminance distribution over a given closed surface surrounding the light source. If the radiant or luminous intensity distribution, or the irradiance or illuminance distribution, is given in polar coordinates (θ, ϕ), the total radiant flux or luminous flux Φ of the light source is given by

$$\Phi = \int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} I(\theta,\phi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{A2.2.13}$$

or

$$\Phi = r^2\int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} E(\theta,\phi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi. \tag{A2.2.14}$$

For example, the total luminous flux of an isotropic point source having luminous intensity of 1 cd is 4π lumens.

Total luminous flux (lumen) is a very important quantity to specify lamp products. It represents how much visible light a lamp can produce (for the given wattage of the lamp), no matter what the intensity distributions are. The ratio of the total luminous flux to the input electrical power (lumens per watt) for a light source—called *luminous efficacy of a light source*—is an important parameter concerned with energy saving, and thus, the importance of accurate measurement of total luminous flux. The total luminous flux of light sources is normally measured either with an integrating sphere photometer (see section A2.2.4.5) or a goniophotometer (see [25]).

A2.2.3.9 Radiance temperature and colour temperature

Radiance temperature (unit: kelvin) is the temperature of the Planckian radiator for which the radiance at the specified wavelength has the same spectral concentration as for the thermal radiator considered. It is commonly used for blackbodies and spectral radiance standard lamps.

Colour temperature (unit: kelvin) is the temperature of a Planckian radiator emitting radiation of the same chromaticity (see section A2.2.6.2) as that of the light source in question. However, actual light sources other than blackbodies rarely have exactly the same chromaticity as a Planckian radiator. Therefore, for various lamps used in general lighting (such as fluorescent lamps and other discharge lamps), another term 'correlated colour temperature' is used. See section A2.2.6.3 for further details.

Distribution temperature (unit: kelvin) is the temperature of a blackbody with a spectral power distribution closest to that of the light source in question, and is used for quasi-Planckian sources such as incandescent lamps. Refer to reference [21] for details.

A2.2.3.10 Relationship between SI units and inch–pound system units

The SI units as described earlier should be used in all radiometric and photometric measurements according to international standards and recommendations on SI units. However, some inch–pound system units are still being used in many application areas. The use of these non-SI units is discouraged. The definitions of such inch–pound system units used in photometry are given in table A2.2.2 for conversion purposes only.

The definition of foot–lambert is such that the luminance of a perfect diffuser is 1 fL when illuminated at 1 fc. In the SI unit, the luminance of a perfect diffuser would be $1/\pi$ ($\mathrm{cd\,m^{-2}}$) when illuminated at 1 lx. For convenience of changing from these inch–pound system units to SI units, the conversion factors are listed in table A2.2.3. For example, 1000 lx is the same illuminance as 92.9 fc, and

Table A2.2.2. Inch–pound system units and their definitions.

foot–candle (fc)	Illuminance	Lumen per square foot ($\mathrm{lm\,ft^{-2}}$)
foot–lambert (fL)	Luminance	$1/\pi$ candela per square foot ($\pi^{-1}\,\mathrm{cd\,ft^{-2}}$)

Note: the use of these non-SI units is discouraged.

Table A2.2.3. Conversion between inch–pound system units and SI units.

To obtain the value in	Multiply the value in	By
lx from fc	fc	10.764
fc from lx	lx	0.09290
$\mathrm{cd\,m^{-2}}$ from fL	fL	3.4263
fL from $\mathrm{cd\,m^{-2}}$	$\mathrm{cd\,m^{-2}}$	0.29186
m (metre) from feet	Feet	0.30480
mm (millimetre) from inch	Inch	25.400

Note: the use of these non-SI units is discouraged.

1000 cd m^{-2} is the same luminance as 291.9 fL. Conversion factors to and from many other units are available in [22].

A2.2.4 Principles in photometry and radiometry

Several important theories in practical photometry and radiometry are introduced in this section.

A2.2.4.1 Inverse square law

Illuminance E (lx) at a distance d (m) from a point source having luminous intensity I (cd) is given by

$$E = \frac{I}{d^2}. \qquad \text{(A2.2.15)}$$

For example, if the luminous intensity of a lamp in a given direction is 1000 cd, the illuminance at 2 m from the lamp in this direction is 250 lx. Or, the luminous intensity I of a lamp is obtained by measurement of illuminance E at distance d from the light source. Note that the inverse square law is valid only when the light source is regarded as a point source. Sufficient distances relative to the size of the source are needed to assume this relationship.

A2.2.4.2 Lambert's cosine law

The luminous intensity of a Lambertian surface element is given by

$$I(\theta) = I_n \cos\theta. \qquad \text{(A2.2.16)}$$

Lambertian surface: A surface whose luminance is the same in all directions of the hemisphere above the surface. The total luminous flux Φ of a Lambertian surface shown in figure A2.2.7 is given by

$$\Phi = \pi I_\mathrm{n} = \pi L a \qquad \text{(A2.2.17)}$$

where L is the luminance of the surface and a the area of the surface.

Perfect (reflecting/transmitting) diffuser: A Lambertian diffuser with a reflectance (transmittance) equal to 1.

A2.2.4.3 Relationship between illuminance and luminance

The luminance L (cd m^{-2}) of a perfect diffuser illuminated by E (lx) is given by (figure A2.2.8)

$$L = \frac{E}{\pi} \qquad \text{(A2.2.18)}$$

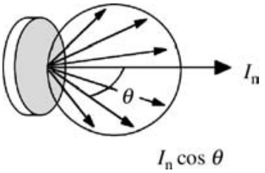


Figure A2.2.7. Lambert's cosine law.

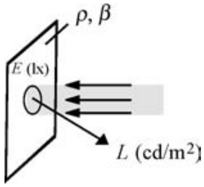


Figure A2.2.8. Relationship between illuminance and luminance.

and, for a Lambertian surface of reflectance ρ,

$$L = \frac{\rho E}{\pi}. \tag{A2.2.19}$$

Reflectance (ρ)

The ratio of the reflected flux to the incident flux in a given condition. The value of ρ can be between 0 and 1.

In the real world, there is neither existing perfect diffuser nor perfectly Lambertian surfaces, and equation (A2.2.19) does not apply for real surfaces. For real object surfaces, the following terms apply.

Luminance factor (β)

Ratio of the luminance of a surface element in a given direction to that of a perfect reflecting or transmitting diffuser under specified conditions of illumination. The value of β can be larger than 1. For a Lambertian surface, reflectance is equal to the luminance factor. Equation (A2.2.19) for a real object is restated using β as

$$L = \frac{\beta E}{\pi}. \tag{A2.2.20}$$

The radiometric term for this quantity is *radiance factor*. Also *reflectance factor* is often used for the same meaning. However, reflectance factor is officially defined for reflected flux in a given cone angle, and it can mean radiance factor (solid angle = 0) or reflectance (solid angle = 2π). See reference [1] for the details.

Note that radiance factor and luminance factor of a real surface vary depending on the illumination and viewing geometry, and thus the geometry must be specified. For example, $\beta_{0/45}$ means radiance (luminance) factor at a 0° incident angle and a 45° viewing angle. See reference [40] for such geometrical notations. Luminance factor also depends on the spectral distribution of the illumination, and thus it must be specified. CIE standard illuminant A or standard illuminant D65 [41] is normally used. Furthermore, radiance factor is given as a function of wavelength.

Luminance coefficient (q)

Quotient of the luminance of a surface element in a given direction by the illuminance on the surface element under specified conditions of illumination.

$$q = \frac{L}{E}. \tag{A2.2.21}$$

Using q, the relationship between luminance and illuminance is thus given by

$$L = qE. \tag{A2.2.22}$$

The radiometric term for this quantity is *radiance coefficient*, used either for light-reflecting or transmitting diffuser materials. The bidirectional reflectance distribution function (BRDF) is also used for the same definition as the radiance coefficient (but expressed as a function of angle of incidence and angle of reflection).

A2.2.4.4 Planck's law

The spectral radiance of a blackbody at a temperature T (K) is given by

$$L_{\mathrm{e}}(\lambda, T) = c_1 n^{-2} \pi^{-1} \lambda^{-5} \left[\exp\left(\frac{c_2}{n\lambda T} \right) - 1 \right]^{-1} \tag{A2.2.23}$$

where $c_1 = 2\pi hc^2 = 3.74177107 \times 10^{-16}\,\mathrm{W\,m^2}$, $c_2 = hc/k = 1.4387752 \times 10^{-2}\,\mathrm{m\,K}$ (1998 CODATA from [23]), h is Planck's constant, c the speed of light in vacuum, k Boltzmann's constant, n the refractive index of medium and λ the wavelength in the medium. $n = 1.00028$ in standard air [6,24].

A2.2.4.5 Principles of integrating sphere

An integrating sphere is a device to achieve spatial integration of luminous flux (or radiant flux) generated (or introduced) in the sphere. In the case of measurement of light sources, the spatial integration is made over the entire solid angle (4π). In figure A2.2.9, assuming that the integrating sphere wall surfaces are perfectly Lambertian, luminance L of a surface element Δa (generated by uniform light incident on this element) creates the equal illuminance E on any part of the sphere surfaces, as given by

$$E = \frac{L\Delta a}{4R^2}. \tag{A2.2.24}$$

In other words, the same amount of flux incident anywhere on the sphere wall will create an equal illuminance on the detector port. In the case of actual integrating spheres, the surface is not perfectly Lambertian, but due to interreflections of light in the sphere, the distribution of reflected light will be sufficiently uniform to approximate the condition assumed in equation (A2.2.24).

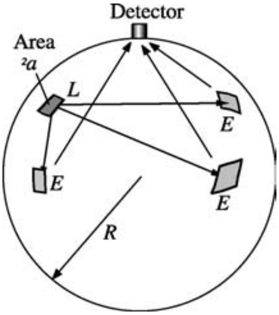


Figure A2.2.9. Flux transfer in an integrating sphere.

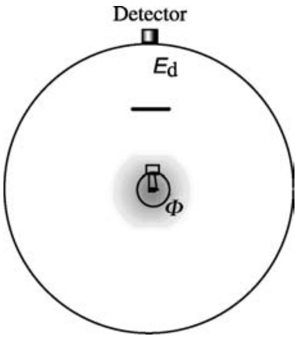


Figure A2.2.10. Integrating sphere photometer.

Integrating sphere photometer

By operating a light source in an integrating sphere as shown in figure A2.2.10, the total luminous flux of the light source is measured using one detector on the sphere wall. Such a device is called an integrating sphere photometer (or Ulbricht sphere). The direct light from an actual light source is normally not uniform, and thus must be shielded from the detector using a baffle. When a light source with luminous flux Φ is operated in a sphere having reflectance ρ, the flux created by interreflections is given by

$$\Phi(\rho + \rho^2 + \rho^3 + \cdots) = \Phi \frac{\rho}{1-\rho}. \tag{A2.2.25}$$

Then, the illuminance E_d created by all the interreflections is given by

$$E_d = \frac{\Phi \rho}{1-\rho} \frac{1}{4\pi R^2}. \tag{A2.2.26}$$

The sphere efficiency (E_d/Φ) is strongly dependent on reflectance ρ due to the term $1-\rho$ in the denominator, where a high reflectance coating such as $\rho = 0.98$ is often used. For this reason, E_d cannot be predicted accurately enough to determine Φ. Real integrating sphere photometers are used as a relative device to measure test lamps against standard lamps whose luminous flux is known. For further details of the integrating sphere photometer, refer to references [25, 26].

A2.2.5 Practice in photometry and radiometry

Photometry and radiometry are practised in many different areas and applications, dealing with various light sources and detectors, and cannot be covered in this chapter. Various references are available on practical measurements in photometry and radiometry. References [26, 27] provide the latest information on standards and practical aspects in photometry. There are a number of publications from CIE on many subjects in photometry including characterization of illuminance meters and luminance meters [28], luminous flux measurement [25], spectroradiometry [29], measurements of LEDs [30], etc. The latest list of CIE publications is available on-line [31]. A series of measurement guide documents are published from the Illuminating Engineering Society of North America (IESNA) for operation and measurement of specific types of lamps [32–34] and luminaires. The American Society for Testing and Materials (ASTM) provides many useful standards and recommendations on optical properties of materials and colour measurements [35]. There are also a number of publications from the National

Institute of Standards and Technology (NIST) on photometry [27], spectral irradiance [36], spectral reflectance [37], spectral responsivity [38], etc.

A2.2.6 Fundamentals of colorimetry

A2.2.6.1 Colour matching functions and tristimulus values

The perception of colour is a psychophysical phenomenon, and the measurement of colour must be defined in such a way that the results correlate accurately with what the visual sensation of colour is to a normal human observer. *Colorimetry* is the measurement science used to quantify and describe physically the human colour perception. The basis of colorimetry was established by CIE in 1931 based on a number of visual experiments that defined a set of three spectral weighting functions [39]. These functions, shown in figure A2.2.11, are called the *CIE 1931 XYZ colour matching functions* denoted as $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$. These functions were derived from a linear transformation of the original set of colour matching functions in such a way that $\bar{y}(\lambda)$ is equal to $V(\lambda)$.

By using the colour matching functions, light stimuli having any spectral power distribution $\phi_\lambda(\lambda)$ can be specified for colour by three values:

$$X = k\int_\lambda \phi_\lambda(\lambda)\bar{x}(\lambda)\,\mathrm{d}\lambda, \quad Y = k\int_\lambda \phi_\lambda(\lambda)\bar{y}(\lambda)\,\mathrm{d}\lambda, \quad Z = k\int_\lambda \phi_\lambda(\lambda)\bar{z}(\lambda)\,\mathrm{d}\lambda \tag{A2.2.27}$$

where $\phi_\lambda(\lambda)$ is the spectral distribution of light stimulus and k a normalizing constant. These integrated values are called *tristimulus values*. Two light stimuli having the same tristimulus values have the same colour even if the spectral distributions are different. For light sources and displays, $\phi_\lambda(\lambda)$ is given in quantities such as spectral irradiance and spectral radiance. If $\phi_\lambda(\lambda)$ is given in an absolute unit (such as $\mathrm{W\,m^{-2}\,nm^{-1}}$, $\mathrm{W\,m^{-2}\,sr^{-1}\,nm^{-1}}$) and $k = 683\,\mathrm{lm\,W^{-1}}$ is chosen, Y yields an absolute photometric quantity such as illuminance (in lux) or luminance (in $\mathrm{cd\,m^{-2}}$). For object colours, $\phi_\lambda(\lambda)$ is given as

$$\phi_\lambda(\lambda) = R(\lambda)S(\lambda) \tag{A2.2.28}$$

where $R(\lambda)$ is the spectral reflectance factor of the object, $S(\lambda)$ the relative spectral distribution of the illumination, and

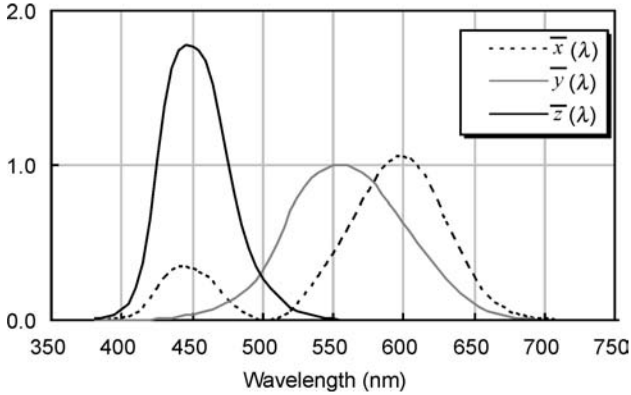


Figure A2.2.11. CIE 1931 XYZ colour matching functions.

$$k = \frac{100}{\int_\lambda S(\lambda)\bar{y}(\lambda)\,\mathrm{d}\lambda} \tag{A2.2.29}$$

so that $Y = 100$ for a perfect diffuser and Y indicates the luminance factor (in %) of the object surface. To calculate colour of objects from spectral reflectance factor $R(\lambda)$, one of the standard illuminants (see section A2.2.6.3) is used.

Tristimulus values can be obtained either by numerical summation of equation (A2.2.27) from the spectral data $\phi_\lambda(\lambda)$ obtained by a spectroradiometer or spectrophotometer, or by broadband measurements using detectors having relative spectral responsivity matched to the colour matching functions. Such a device using three (or four) detector channels is called *tristimulus colorimeter*.

When applying colorimetric data for real visual colour matching, it should be noted that the $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ colour matching functions are based on experiments using 2° field of view and applicable only to narrow fields of view (up to 4°). Such an ideal observer is called the *CIE 1931 standard colorimetric observer*. In 1964, the CIE defined a second set of standard colour matching functions for a 10° field of view, denoted as $\bar{x}_{10}(\lambda)$, $\bar{y}_{10}(\lambda)$, $\bar{z}_{10}(\lambda)$, to supplement those of the 1931 standard observer. This is called the *CIE 1964 supplementary standard colorimetric observer*, and can be used for a field of view greater than 4°. The 2° observer is used in most applications for colorimetry of light sources. The 10° observer is often used in object colour measurements. For further details of colorimetry and colour science, refer to official CIE publications [40–42] and many other general references [43].

A2.2.6.2 Chromaticity diagrams

While the tristimulus values can specify colour, it is difficult to associate what colour it is from the three numbers. By projecting the tristimulus values onto a unit plane ($X + Y + Z = 1$), colour of light can be expressed on a two-dimensional plane. Such a unit plane is known as the *chromaticity diagram*. The colour can be specified by the *chromaticity coordinates* (x,y) defined by

$$x = \frac{X}{X+Y+Z}, \qquad y = \frac{Y}{X+Y+Z}. \tag{A2.2.30}$$

The diagram using the chromaticity coordinates (x,y), as shown in figure A2.2.12(a), is referred to as the *CIE 1931 chromaticity diagram*, or the *CIE (x,y) chromaticity diagram*. The boundaries of this horseshoe-shaped diagram are the plots of monochromatic radiation (called the *spectrum locus)*.

The (x,y) chromaticity diagram is significantly nonuniform in terms of colour difference. The minimum perceivable colour differences in the CIE (x,y) diagram, known as the *MacAdam ellipses*, are shown in figure A2.2.12(a). To improve this, in 1960, CIE defined an improved diagram—*CIE 1960 (u,v) chromaticity diagram* (now obsolete), and in 1976, a further improved diagram—*CIE 1976 uniform chromaticity scale (UCS) diagram*, as shown in figure A2.2.12(b), with its chromaticity coordinate (u',v') given by

$$u' = \frac{4X}{X+15Y+3Z}, \qquad v' = \frac{9Y}{X+15Y+3Z}. \tag{A2.2.31}$$

While the (u',v') chromaticity diagram is a significant improvement from the (x,y) diagram, it is still not satisfactorily uniform. Both of these diagrams are widely used. Note that these chromaticity diagrams are intended to present colour of light sources (emitted light) and not colour of objects (reflected light). Presentation of object colours requires a three-dimensional colour space that incorporates another dimension—lightness (black to white). Refer to other publications [40, 43] for the details of object colour specification.

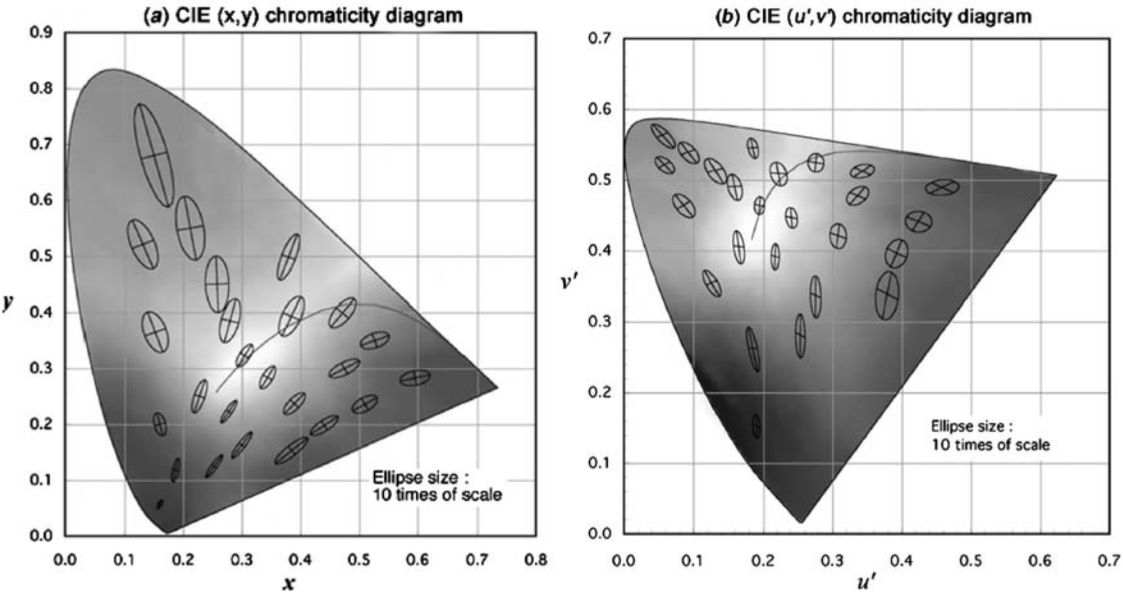


Figure A2.2.12. (*a*) MacAdam ellipses on CIE 1931(x,y) diagram and (*b*) MacAdam ellipses on the CIE 1976 ($u'v'$) diagram. The ellipses are plotted 10 times their actual size. The curve near the centre region is the Planckian locus.

A2.2.6.3 Colour temperature and correlated colour temperature

Figure A2.2.13 shows the trace of the (x,y) chromaticity coordinate of blackbody radiation (see section A2.2.4.4) at its temperature from 1600 to 20 000 K. This trace is called the *Planckian locus*. The colours on the Planckian locus can be specified by the blackbody temperature in kelvin and is called colour *temperature* (see also section A2.2.3.9). The colours around the Planckian locus from about 2500 to 20 000 K can be regarded as *white*, 2500 K being reddish white and 20 000 K being bluish white. The

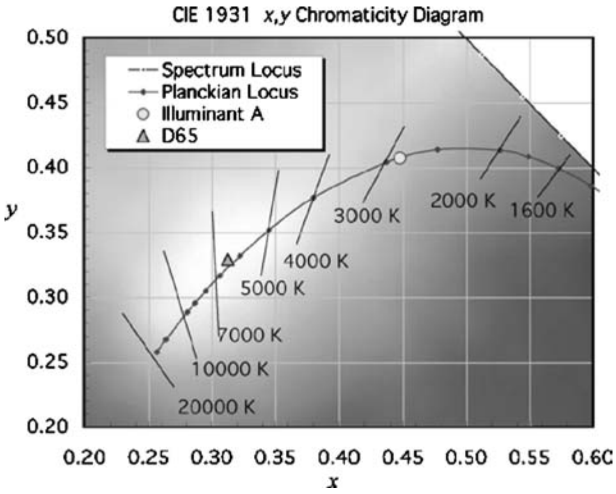


Figure A2.2.13. Planckian locus on (x,y) chromaticity diagram and iso-CCT lines.

point labelled 'Illuminant A' is the typical colour of an incandescent lamp, and 'Illuminant D65' the typical colour of day light, as standardized by the CIE [41]. The colours of most traditional lamps for general lighting fall in the region between these two points (2800–6500 K). Strictly speaking, colour temperature cannot be used for colours away from the Planckian locus, in which case *correlated colour temperature* (CCT) is used. CCT is the temperature of the blackbody whose perceived colour most closely resembles that of the light source in question [2]. Due to the nonuniformity of the x, y diagram, the iso-CCT lines are not perpendicular to the Planckian locus on the x, y diagram (see figure A2.2.13). To calculate CCT, therefore, one of the improved uniform chromaticity diagrams is used. Due to the long tradition, CIE specifies that the 1960 (u,v) diagram (now obsolete for other purposes) be used, where the iso-CCT lines are perpendicular to the Planckian locus by definition. From (u',v') coordinates, (u,v) can be obtained by $u = u'$, $v = 2v'/3$. On the (u,v) diagram, find the point on the Planckian locus that is at the shortest distance from the given chromaticity point. CCT is the temperature of the Planck's radiation at that point. A practical way of computing CCT is available [44].

A2.2.6.4 Colour rendering index

For light sources for lighting applications, it is important to evaluate how well their illumination can render colours of objects. The CIE defines the *colour rendering index* (CRI) [45]. The CRI is calculated from the spectral distribution of light under test and the spectral reflectance factor data of 14 Munsell colour samples. The colour difference ΔE_i (on the 1964 $W^*U^*V^*$ uniform colour space—now obsolete) of each sample illuminated by the light under test and by a reference source (Planckian radiation for CCT < 5000 K or a daylight illuminant for CCT ≥ 5000 K) is calculated taking into account chromatic adaptation. The *special colour rendering index* R_i for each colour sample is calculated by

$$R_i = 100 - 4.6\Delta E_i. \qquad \text{(A2.2.32)}$$

This gives an indication of colour rendering for each particular colour. The *general colour rendering index*, R_a, is given as the average of the first eight colour samples (medium saturation). With the maximum value being 100, R_a gives a scale that expresses well the visual impression of colour rendering. For example, lamps having R_a values greater than 80 may be considered suitable for interior lighting, and R_a greater than 95 for visual inspection purposes, etc. See reference [45] for further details.

A2.2.6.5 Colour quantities for LEDs

In addition to chromaticity coordinates x, y and u', v', the following quantities are used to specify the colour and spectrum of LEDs. The definitions in this section follow reference [30].

Peak wavelength λ_p: The wavelength at the maximum of the spectral distribution.

Spectral bandwidth (at half intensity level) $\Delta\lambda_{0.5}$: Calculated as the width between the wavelengths at half of the peak of spectral distribution, as shown in figure A2.2.14. It is also denoted as $\Delta\lambda$(FWHM).

Centroid wavelength λ_c: Calculated as the 'centre of gravity wavelength', according to the equation

$$\lambda_c = \frac{\int_\lambda \lambda S(\lambda)\, d\lambda}{\int_\lambda S(\lambda)\, d\lambda} \qquad \text{(A2.2.33)}$$

Dominant wavelength λ_d: Wavelength of the monochromatic stimulus that, when additively mixed in suitable proportions with the specified achromatic stimulus, matches the colour stimulus considered. Equal energy spectrum with $(x, y) = (0.3333, 0.3333)$ is used as the achromatic stimulus. See figure A2.2.15, where N denotes the achromatic stimulus.

Exitation purity p_e: Defined as the ratio NC/ND in figure A2.2.15. The value of excitation purity is unity if the chromaticity of the LED is on the spectrum locus.

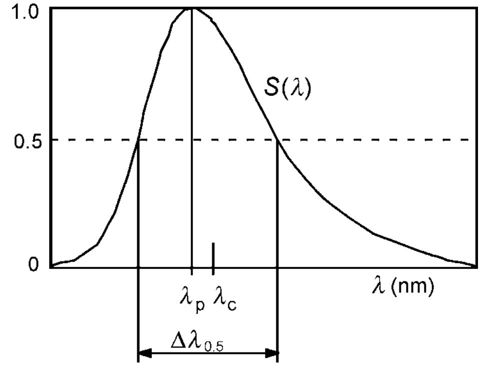


Figure A2.2.14. Typical relative spectral distribution of an LED.

A2.2.6.6 Spectroradiometry for LED colour measurement

Spectroradiometers are commonly used for colour measurement of LEDs. Tristimulus colorimeters are rarely used for LED colour measurements because errors tend to be too large for measurement of quasi-monochromatic sources such as LEDs. Major sources of error in spectroradiometric measurements of LEDs are bandwidth, wavelength error and stray light. To obtain the colour quantities of LEDs accurately, a monochromator bandwidth (FWHM) of 5 nm or less is recommended. The scanning interval should be matched with the bandwidth, or much smaller than the bandwidth. If the bandwidth of the instrument is larger than 5 nm, the bandwidth error can be very large. The errors with a 10 nm bandwidth can be up to 0.005 in x, y depending on the LED peak wavelength [46]. A method for correcting the bandwidth error is available in the case where the bandwidth is a triangular shape and

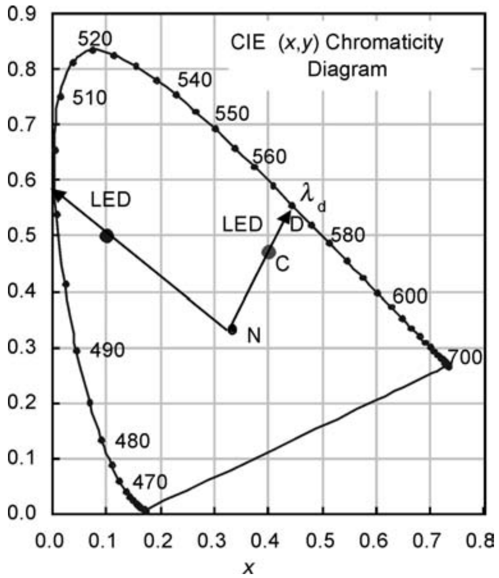


Figure A2.2.15. (x,y) chromaticity diagram showing the dominant wavelength and excitation purity.

is matched with the scanning interval [47]. The wavelength accuracy is also critical for LED colour measurements. An error for 1 nm in wavelength scale would lead to a maximum error of 0.02 in x, y, depending on the LED peak wavelength. The stray light of a monochromator is not very critical for colour measurement of broadband white light sources such as fluorescent lamps, but it is critical for LEDs having narrowband emissions. The errors should be examined for single monochromators including diode-array spectroradiometers. For the details of spectroradiometry in general, refer to other references [29, 48].

References

[1] *International Lighting Vocabulary* 1987 (CIE Publication No.17.4 /IEC Publication 50 (845))

[2] Ohno Y 2001 Chapter 14, Photometry and radiometry—review for vision optics, Part 2 vision optics *OSA Handbook of Optics* vol 3 (New York: McGraw-Hill)

[3] *International Vocabulary of Basic and General Terms in Metrology* 1994 (BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML)

[4] *Quantities and Units: ISO Standards Handbook* 1993 3rd edn

[5] CIE Compte Rendu 1924 p 67

[6] *The Basis of Physical Photometry* 1983 (CIE Publication No.18.2)

[7] *Principles Governing Photometry* 1983 (Sèvres, France: Bureau International Des Poids et Mesures (BIPM) Monograph, BIPM)

[8] *CIE Disk D001 Photometric and Colorimetric Tables* 1988

[9] *CGPM, Comptes Rendus des Séances de la 16e Conférence Générale des Poids et Mesures* 1979 (Sèvres, France: Paris 1979, BIPM)

[10] *CIPM, Comité Consultatif de Photométrie et Radiométrie 10e Session–1982* 1982 (Sèvres, France: BIPM, Pavillon de Breteuil)

[11] *CIE Compte Rendu* 1951 vol 3, Table II, pp 37–39

[12] *CIPM Procés-Verbaux 44* 1976 p 4

[13] *Mesopic Photometry: History, Special Problems and Practical Solutions* 1989 (CIE Publication No. 81)

[14] Walsh J W T 1953 *Photometry* (London: Constable)

[15] Blevin W R and Steiner B 1975 *Metrologia* **11** 97

[16] Blevin W R 1979 The candela and the watt *CIE Proc.* P-79-02

[17] *Le Système International d'Unité (SI), The International System of Units (SI)* 1991 6th edn Sèvres, France: Bur. Intl. Poids et Mesures

[18] Taylor B N 1995 *Guide for the Use of the International System of Units (SI)* (Natl. Inst. Stand. Technol. Spec. Publ. 811)

[19] Taylor B N, ed 1991 *Interpretation of the SI for the United States and Metric Conversion Policy for Federal Agencies* (Natl. Inst. Stand. Technol. Spec. Publ. 814)

[20] *SI Units and Recommendations for the Use of Their Multiples and of Certain Other Units* 1992 (ISO 1000:1992, International Organization for Standardization, Geneva, Switzerland)

[21] *CIE Collection in Photometry and Radiometry* 1994 (Publication No.114/4)

[22] *Appendix, the Lighting Handbook 9th Edition, Illuminating Engineering Society of North America* 2000

[23] *J. Phys. and Chem. Ref. Data* 1999 **28** 1713–1852

[24] Blevin W R 1972 Corrections in optical pyrometry and photometry for the refractive index of air *Metrologia* **8** 146

[25] *CIE Publication No. 84* 1989 Measurements of Luminous Flux

[26] C DeCusatis, ed 1997 *OSA/AIP Handbook of Applied Photometry* (Woodbury, NY: AIP Press)

[27] Ohno Y 1997 *Photometric Calibrations*, NIST Special Publication 250-37. This document is available at http://physics.nist.gov/photometry

[28] *Methods of Characterizing Illuminance Meters and Luminance Meters* 1987 (CIE Publication 69)

[29] *The Spectroradiometric Measurement of Light Sources* 1984 (CIE Publication 63)

[30] *Measurement of LEDs* 1997 (CIE Publication 127)

[31] CIE Central Bureau website: http://www.cie.co.at/cie

[32] *IES Approved Method for the Electric and Photometric Measurement of Fluorescent Lamps* 1999 (IESNA LM-9)

[33] *Electrical and Photometric Measurements of General Service Incandescent Filament Lamps* 2000 (IESNA LM-45)

[34] *Electrical and Photometric Measurements of Compact Fluorescent Lamps* 2000 (IESNA LM-66)

[35] *ASTM Standards on Colour and Appearance Measurement* 1996 5th edn

[36] Walker J H, Saunders R D, Jackson J K and McSparron D A 1987 *Spectral Irradiance Calibrations* (NBS Special Publication 250-20)

[37] Barnes P Y, Early E A and Parr A C 1998 *Spectral Reflectance* (NIST Special Publication 250-41)

[38] Larason T C, Bruce S S and Parr A C 1998 *Spectroradiometric Detector Measurements* (NIST Special Publication 250-48)

[39] *CIE Compte Rendu* 1931 Table II, pp 25–26

[40] CIE Publ. No.15.2 1986 *Colorimetry* 2nd edn
[41] ISO10526/CIE5005: CIE standard illuminants for colorimetry 1999
[42] ISO/CIE10527-1991, *CIE standard colorimetric observers* 1991
[43] For example, Wyszecki G and Stiles W S 1982 *Colour Science: Concepts and Methods, Quantitative Data and Formulae* (New York: Wiley)
[44] Robertson R 1968 Computation of correlated colour temperature and distribution temperature *J. Opt. Soc. Am.* **58** 1528–1535
[45] *Method of Measuring and Specifying Colour Rendering Properties of Light Sources* 1995 (CIE Publ. No. 13.3)
[46] Jones C F and Ohno Y 1999 Colorimetric Accuracies and Concerns in Spectroradiometry of LEDs *Proc., CIE Symposium'99—75 Years of CIE Photometry, Budapest* 173–177
[47] Stearns E and Stearns R 1988 An example of a method for correcting radiance data for bandpass error, Color Res. Application **13-4** 257–259
[48] For example, Kostkowski H J 1997 *Reliable Spectroradiometry* (La Plata, MD: Spectroradiometry Consulting)

A2.3
Nonlinear and short pulse effects

Günter Steinmeyer

A2.3.1 Introduction

In most parts of this handbook, we deal with *linear optical effects*. Linear optics means that the optical power at the outputs of an optical device always scales linearly with input power. The device may spectrally or spatially filter the input beam; it may split the input beam into a multitude of output beams; regardless of what the device does, the output power always relates linearly to the input power. Looking through textbooks on classical optics from the pre-laser era, the impression may arise that the linearity of optical phenomena is a given thing as there is no mention of any *nonlinear effects*. This is in strong contrast, e.g. to acoustics, where nonlinearities are so widespread that the art lies more in their avoidance than in the observation of nonlinearities. One may think of a cheap set of speakers just as one simple example. Increasing the volume, these speakers will start to sound increasingly annoying with more and more audible distortion. This distortion is not related to the frequency-dependence of the speaker's transmission characteristics, because this would be independent of volume. The distortion effect is related to unwanted *harmonics* of the input. These harmonics arise due to nonlinearities between the emitted acoustic wave and the input current to the speaker's solenoid. Beyond a certain drive amplitude, the speaker's membrane position does not linearly follow the current any more. These harmonics are not necessarily bad. All musical instruments also rely on acoustic nonlinearities, giving rise to a characteristic spectrum of overtones of the excited fundamental vibration of a chord. This characteristic spectrum allows us to distinguish different musical instruments. The omnipresence of nonlinearities in acoustics is in strong contrast to optics, where similar effects could not be observed until the advent of the laser.

In a nonlinear system, the output power has to scale nonlinearly with input power. One such effect in optics, which is similar to the harmonics in acoustics, is called *second-harmonic generation* (*SHG*). In this nonlinear optical effect, the *overtone* of an optical wave is generated inside a crystal, e.g. converting the infrared light of a laser into bright green light with half the wavelength of the input light. The output intensity of the green light scales quadratically with the input power, which is a clear indication of a nonlinear optical process of second order. SHG is an example of a *degenerate* nonlinear optical process, i.e. all interacting waves carry the same wavelength. In a crystal with such a second-order nonlinearity, we may also observe *non-degenerate* processes. Non-degeneracy means *mixing* of two input waves with different wavelength, generating the *sum-frequency* or the *difference-frequency* of the two input waves. Now the intensity of the generated sum-frequency light scales with the product of the two input intensities. Note that all nonlinearities discussed so far only involve optical fields. That means that they are *all-optical nonlinear effects*. This has to be seen in contrast to nonlinearities involving one low-frequency electric field (as accessible by conventional electronics) and one optical field. This interaction may be thought of as a limiting case of sum-frequency generation (SFG) with one of the fields close to zero frequency. This mixing effect can be used for building *electro-optical modulators*. However, the *optical* output power of the device scales *linearly* with the *optical* input power. Therefore, electro-optic mixing is not considered a nonlinear optical effect. The same argument holds for *acousto-optics*, where there is in fact mixing between acoustic waves and the optical field. Clearly, acousto-optics is also not considered a nonlinear optical effect, as power scaling is entirely linear for the optical field involved. In the case of the electro-optic effect, we note that this distinction may appear somewhat arbitrary looking at the underlying physics of the process. This distinction will become much clearer from the different regime of applications of all-optical nonlinearities compared to optical modulators.

With the high intensities of laser sources, an all-optical nonlinear regime can easily be reached. The higher the peak power, the more pronounced the nonlinear behaviour of certain optical materials. This makes it clear that a pulsed laser source generates much more abundant nonlinearities compared to a continuous wave source, which is why we refer to nonlinear processes also as *short pulse effects*. In the high-peak-power regime the response of optical materials starts to react nonlinearly, both in amplitude and phase, with incident optical power. A further simple example for nonlinear behaviour may be the bleaching of an optical neutral density filter in a pulsed high-power laser beam. As soon as a certain intensity is surpassed, the attenuation of the filter starts to decrease, because a substantial part of the carriers within the beam cross section is transferred into the conduction band. In turn, the density of carriers in the valence band is decreased. These *carrier-related* effects therefore cause a *saturation* of the absorption in this material.

In optoelectronics, such nonlinear optical behaviour is particularly interesting for the construction of an *all-optical switch*, a device that allows controlling light with light. We will use the term *optical switch* as a general term for very different applications of optical nonlinearities, including optical gates, mode-lockers, and optical limiters. For some applications, one may think of an optical switch acting like an *optical transistor* or *light valve*, employing the nonlinearity for modulation of one light beam by another one. Other applications rely on the *self-action* of intense short pulses in a nonlinear medium. Here the pulse modulates its own amplitude or phase profile via a nonlinear optical effect. Consequently, one talks of *self-amplitude modulation (SAM)* or *self-phase modulation (SPM)*, respectively. Quite generally all methods of all-optical switching outperform acousto-optic or electro-optic concepts in terms of speed. On the other hand, it is very challenging for an all-optical switch to reach anything close to the contrast and efficiency possible with electro-optic switches. This trade-off between speed and switching contrast appears to be the fundamental dilemma of all types of optical switches. In this chapter, we will first give a general overview of nonlinear optical processes. Because of their importance in optoelectronics, we will then concentrate on applications of nonlinear optical processes as all-optical switches.

A2.3.2 Quasi-instantaneous nonlinear optical processes

In the introduction, we have already given examples for the two main types of nonlinear processes. Saturable absorption due to *carrier transfer* into the excited state is a process with *limited response times*, which typically lie in the *picosecond range*. *Non-resonant nonlinearities* such as SHG occur in totally *transparent dielectric* materials and can be significantly faster. As no direct transfer of carriers into any kind of excited state takes place, there is also no limitation by relaxation processes. In the non-resonant case *bound electrons* contribute to the nonlinearity instead. One can coarsely estimate the order of magnitude of the response time τ of nonlinearities in dielectric materials as $\tau = 1/\Delta\nu$, where $\Delta\nu$ is the transparency frequency range of the material. This immediately brings us to single femtoseconds response times for materials that are transparent in the visible [1]. We will therefore refer to these as materials with *quasi-instantaneous* response. For simplicity we will often relate to these nonlinear processes as *instantaneous* effects. Instantaneous effects comprise sum and difference frequency generation, parametric interaction, and also phase effects such as SPM, see figure A2.3.1.

Given the instantaneous nature of the electronic nonlinearities, one can directly write the polarization $\vec{P}$ in a dielectric material as a function of the electric field $\vec{E}$ [1, 2]

$$P_i = \sum_{j=1}^{3} \chi_{ij}^{(1)} E_j + \sum_{j,k=1}^{3} \chi_{ijk}^{(2)} E_j E_k + \sum_{j,k,l=1}^{3} \chi_{ijkl}^{(3)} E_j E_k E_l + \dots \tag{A2.3.1}$$

The susceptibility coefficients $\chi^{(i)}$ are tensors of corresponding rank $i + 1$. The linear optical properties of a material are all included in the first order coefficient $\chi^{(1)}$. In many materials, all tensor elements $\chi_{ij}^{(1)}$ will be identical and we can use simple scalar susceptibility coefficients $\chi^{(1)}$ instead. Only materials exhibiting birefringence or dichroism require a full-blown tensorial treatment of the susceptibility.

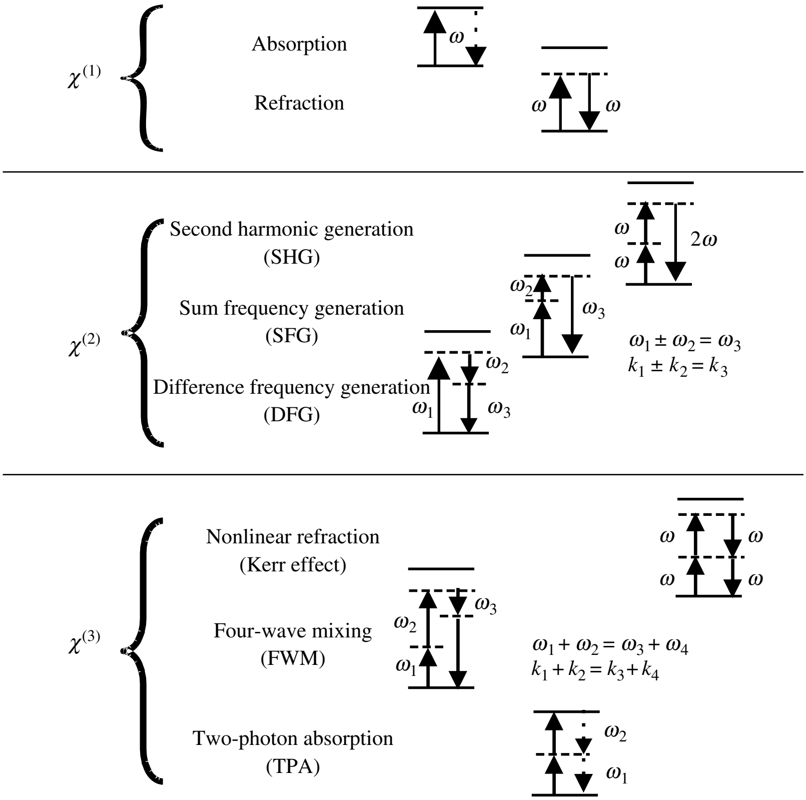

$\chi^{(1)}$
Absorption
Refraction
ω
ω
ω
$\chi^{(2)}$
Second harmonic generation (SHG)
Sum frequency generation (SFG)
Difference frequency generation (DFG)
ω
ω
2ω
ω_2
ω_1
ω_3
ω_2
ω_1
ω_3
$\omega_1 \pm \omega_2 = \omega_3$
$k_1 \pm k_2 = k_3$
$\chi^{(3)}$
Nonlinear refraction (Kerr effect)
Four-wave mixing (FWM)
Two-photon absorption (TPA)
ω
ω
ω
ω
ω_3
ω_2
ω_1
$\omega_1 + \omega_2 = \omega_3 + \omega_4$
$k_1 + k_2 = k_3 + k_4$
ω_2
ω_1

Figure A2.3.1. Linear and nonlinear optical processes in the photon picture. Real states are designated by solid horizontal lines; virtual states by dashed lines. Linear absorption is caused by a transition between two real states with the potential relaxation indicated by a dotted arrow, refraction is caused by absorption and delayed re-emission of a photon by a virtual state. In a virtual process, the energy of all photons is conserved. In the virtual $\chi^{(2)}$ processes, three waves interact in different sum and difference frequency generation schemes. In this picture, the $\chi^{(3)}$ process of nonlinear refraction (all-optical Kerr effect) can also be considered as the degenerate case of the more general four-wave mixing process. $\chi^{(2)}$ processes can only occur in anisotropic materials. $\chi^{(3)}$ processes occur in any kind of medium and are of particular interest in fibre optics. Two-photon absorption ends in an excited state. The potential return path is again indicated by a dotted arrow.

Otherwise, we can use a simplified version of equation (A.2.3.1)

$$\vec{P} = \chi^{(1)}\vec{E} + (\chi^{(2)}|\vec{E}|)\vec{E} + (\chi^{(3)}|\vec{E}|^2)\vec{E} + \ldots. \tag{A2.3.2}$$

The complex index of refraction is related to the susceptibility via $n + i\kappa = \sqrt{1 + \chi^{(1)}}$, i.e. $\chi^{(1)}$ contains information about both the refractive index n and the absorption coefficient κ. Nonlinear optical processes are included in the higher-order coefficients and again may affect both the amplitude and the phase of the light. In isotropic materials such as glasses all tensor elements are degenerate and the susceptibility coefficients of even order vanish, i.e. $\chi^{(2)} = \chi^{(4)} = \ldots = 0$. Anisotropy can be found in many crystals and is often, but not necessarily, accompanied by birefringence. Tabulated second-order coefficients $\chi^{(2)}$ for nearly all relevant second order materials can be found in [2].

In figure A2.3.1, some of the most commonly encountered optical nonlinearities up to order three are summarized. The linear processes of absorption and refraction are also included for comparison. For most of these nonlinear optical processes, $\sum_i \pm \hbar\omega_i = 0$ has to hold as a consequence of energy conservation, see figure A2.3.1. This is not true for absorption and two-photon absorption (TPA), where energy is stored in an excited state. Here $\sum_i \pm \hbar\omega_i = E_{\mathrm{abs}}$ has to be fulfilled instead. This means that the sum of all photon energies equals the energy of the excited state E_{abs} in the case of one or multiphoton absorption. If only virtual states are involved, no energy is absorbed in the process and the sum has to vanish. The latter is the case for *SHG*, *difference frequency generation* (*DFG*), and *SFG*. Some of the $\chi^{(3)}$ processes shown in figure A2.3.1 also only involve virtual excited states. In all these non-absorptive cases, momentum conservation $\sum \pm k_i = 0$ also has to hold, where the wavenumber k_i is defined as $\omega_i n/c$. One may think of the real processes absorption and two-photon absorption as frustrated refraction or four-wave mixing processes, respectively. These frustrated processes end in a real rather than virtual state and can therefore not relax back to the ground state.

It is useful to relate equation (A2.3.2) to our initial acoustics example, where the electric field $\vec{E}$ takes the role of the solenoid drive current and the polarization $\vec{P}$ is related to the membrane position. As soon as a significant contribution from the nonlinear terms in equation (A2.3.2) appears, we would expect to see optical harmonics in analogy to acoustics. Let us come back to our original example for nonlinear optical behaviour, i.e. SHG. This effect manifests itself as a parabolic susceptibility term $\chi^{(2)}$ in equation (A2.3.2). In such a medium, the polarization P does not proportionally follow the electric field $E \propto \sin(\omega t)$ of the input field. As illustrated in figure A2.3.2, the resulting polarization

$$P \propto \chi^{(1)}\sin(\omega t) + \chi^{(2)}\sin^2(\omega t) = \chi^{(1)}\sin(\omega t) + \frac{\chi^{(2)}}{2}[1 - \cos(2\omega t)] \tag{A2.3.3}$$

clearly exhibits distortions compared to the sinusoidal input field, which manifest themselves as *frequency-doubled components* in the polarization [1, 3]. The output field therefore consists of the fundamental and its second harmonic (and a constant term, which does not propagate). If we had investigated the $\chi^{(3)}$ term instead, we would find the third harmonic of the input field. Similarly, one can easily show that higher-order terms $\chi^{(i)}$ lead to the generation of the i-th harmonic in the polarization. The coefficients $\chi^{(i)}$ therefore govern the characteristic spectrum of harmonics of a nonlinear material, similar to the relative strength of overtones in acoustics discussed in the introduction. In the non-degenerate case of mixing of two electric fields of different frequency, one observes both the sum and difference frequency mixing terms in the output spectrum.

A2.3.3 Scattering processes—coupling to the lattice

The effects described so far only encompassed interaction of photons and electrons. A third category of nonlinear optical processes comprises scattering processes, in particular those involving phonons in

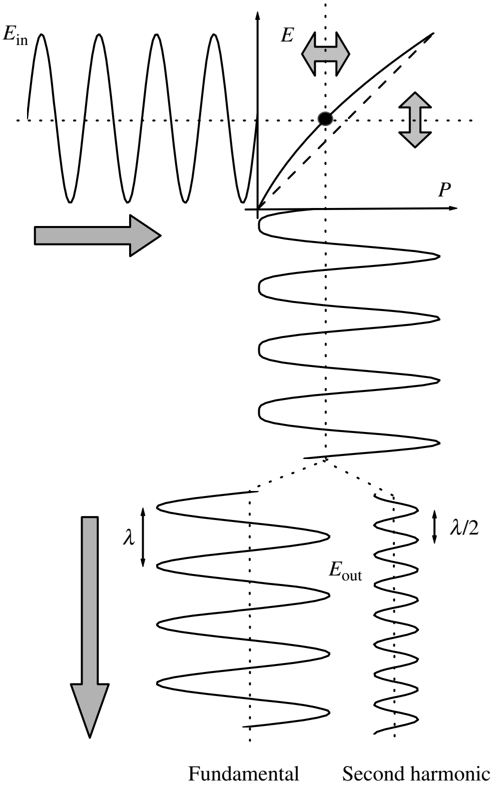


Figure A2.3.2. Generation of the second harmonic. If the polarization P inside a medium does not simply proportionally follow the input field E_{in} but $P(E)$ includes a parabolic term, the resulting polarization contains non-sinusoidal components, which reveal themselves as a harmonic contribution at double the frequency of the input field. Generally, a $\chi^{(i)}$ susceptibility term gives rise to generation of the i-th harmonic in the output field.

solid-state materials. Coupling of lattice vibrations and the electric field give rise to *Brillouin scattering* (acoustic phonons) and *Raman scattering* (optical phonons) [3, 4]. These processes can be understood similar to the DFG and SFG processes but with one photon replaced by a quantized lattice vibration, i.e. a phonon. Again, these processes involve virtual intermediate states, see figure A2.3.1. The decay of one input photon into a photon of smaller energy and one phonon is referred to as the Stokes process. The wavelength shift accompanying this scattering process is called the *Stokes-shift*. Therefore, looking at scattering processes from a photon point of view, energy conservation does not appear to hold as the excess energy is transferred into the phonon. The sum of all particle energies, however, is conserved. At high light intensities, the opposite process similar to the sum-frequency process may also show up. In this case, a previously generated phonon may add to the energy of one photon, with a resulting shorter-wavelength photon as the product of this so-called anti-Stokes process.

The scattering processes are difficult to categorize, as they rely on the same kind of mechanism as SFG but are linear in terms of optical power. Therefore, they do not qualify as nonlinear optical processes similar to the electro-optic processes discussed in the introduction. Nevertheless, the appearance of Raman and Brillouin scattering is closely related to nonlinear processes as are the electro-optic processes.

At high pump intensities a new phenomenon called stimulated scattering can be observed. Here an exponential increase of the Stokes wave with input power is observed. This process is nonlinear, whereas the spontaneous scattering described above has to be considered a linear optical process. These processes have found many applications in fibre optics, e.g. for Raman lasers and frequency conversion. For a detailed description of stimulated scattering processes and their applications we refer the reader to specialized literature on nonlinear fibre optics [4].

A2.3.4 Saturable absorbers

A2.3.4.1 Types of saturable absorbers

So far, we have mainly addressed processes which follow the polarization in the medium instantaneously. Many nonlinear processes, in particular those involving *generation of free carriers*, exhibit a non-instantaneous response. For continuous light exposure or at long pulse durations, the nonlinearity of these processes can also be expressed as a perturbative series like equation (A2.3.2). On shorter time scales, however, these processes may not relax quickly enough, and the number of free carriers grows linearly with the accumulated exposure. As the number of carriers is responsible for the nonlinearity, there is also a cumulative effect for the nonlinearity of the process. For short pulses (short meaning faster than the relaxation time constant), the nonlinearity therefore rather scales with pulse energy than with peak power.

Saturable absorption is the prime example of such an effect. It is also a very useful effect for constructing nonlinear optoelectronic devices. A *saturable absorber* acts like a switch that allows, for example, discrimination of low-energy pulses inside a laser. As short pulses with sufficiently high energy experience a decreased optical loss by the saturable absorption, a saturable absorber effectively favours pulsed operation of the laser over continuous operation. Such devices typically contain one or several saturable absorbers and a reflector. These devices have been named *saturable Bragg reflector* (SBR, [5, 6]) or *semiconductor saturable absorber mirror* (SESAM, [7]) or sometimes simply *saturable absorber mirror*. As these optoelectronic devices have become an important element in the construction of mode-locked lasers, we will outline some fundamental physical processes contributing to the nonlinearity of such devices. We will also give an overview on the construction of these optoelectronic building blocks.

In principle, any absorbing material could be used to build a saturable absorber. Dyes have been very popular for that purpose in the 1970s and 1980s [8, 9]. Dyes offer some adaptability by changing concentration and the centre wavelength of the absorber. Several solid-state materials based on doped glasses have been proposed as an alternative [10, 11]. As a crystalline material, Cr:YAG has found widespread applications as an intracavity switch [12–15]. The potential of semiconductor materials for electro-optical switches, in particular that of quantum well absorbers, has already been recognized in the 1970s and 1980s [16–20]. As the optical properties of quantum wells can be tailored in the most flexible way of all materials discussed so far [21, 22], this type of saturable absorber has become the method of choice for all-optical switches and has found widespread use.

A2.3.4.2 The physics of saturable absorption

The fundamental processes involved in saturable absorption are schematically illustrated in figure A2.3.3. Initially, a short laser pulse transfers electrons from the valence band of the absorber into the conduction band of the material. Again, short means that there is only insignificant relaxation back to the valence band within the pulse duration and some time after. By depleting the valence band this process therefore reduces the density of carriers that are capable of absorbing the laser light, which effectively decreases absorption [21]. A short powerful laser pulse can generate a *nonthermal carrier*

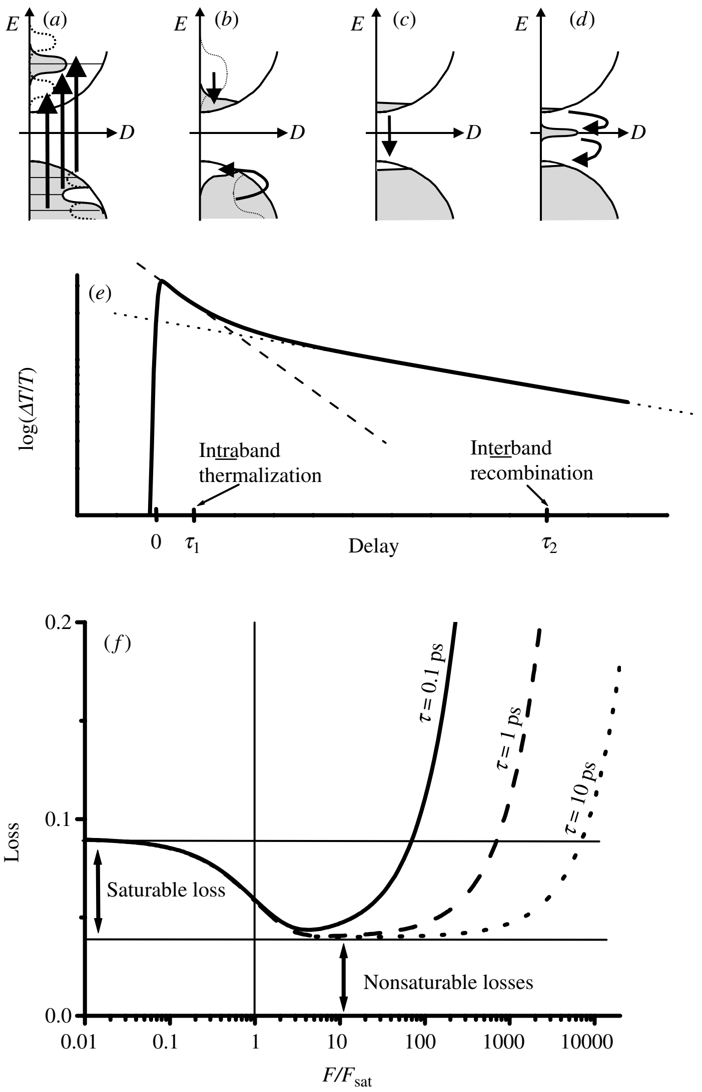

E
(a)
D
E
(b)
D
E
(c)
D
E
(d)
D
(e)
log(ΔT/T)
Intraband thermalization
Interband recombination
0
τ1
Delay
τ2
(f)
0.2
0.1
0.0
Loss
τ = 0.1 ps
τ = 1 ps
τ = 10 ps
Saturable loss
Nonsaturable losses
0.01
0.1
1
10
100
1000
10000
F/Fsat

Figure A2.3.3. A detailed look at saturable absorption. After excitation of carriers from the valence band of a semiconductor into its conductance band (*a*), the initial non-thermal distribution tends to rapidly cool (*b*) with a time constant τ_1. On this time scale of int*ra*band relaxation, the majority of carriers is still in the excited state and only relaxes into the valence band on longer time scale, which is called int*er*band relaxation (*c*). Introducing additional mid-gap states as shown in (*d*) may greatly accelerate the interband recombination. Monitoring the induced transmission change $\Delta T/T$ as shown in (*e*) identifies intraband and interband relaxation by two distinct time constants, τ_1 and τ_2. (*f*) shows the typical dependence of the induced increase of transmission as a function of input fluence. For input fluence well below the saturation fluence F_{sat}, the overall loss of the device amounts to the sum of saturable and non-saturable losses. At higher fluences, the loss starts to roll off in the vicinity of F_{sat} and reaches a minimum value dictated by non-saturable losses L_{NS}. At still higher fluence, one may observe an increased loss due to two-photon processes.

distribution high up in the valence band [22, 23] as schematically shown in figure A2.3.3(*b*). These hot electrons are not in thermal equilibrium with the lattice and tend to rapidly cool via phonon scattering processes. Typically, this *intraband relaxation process* occurs on a time scale of $\tau_1 \approx 100\,\mathrm{fs}$ (figure A2.3.3(*c*)). The cooled carriers will fill the bottom of the conduction band, relaxing to the valence band on much longer time scales τ_2, which are typically of the order of nanoseconds. This is also called *interband relaxation*. The two different timescales of carrier cooling and recombination generate a marked *bitemporal behaviour* [7] of the absorption *vs*. time, as shown in figure A2.3.3(*e*).

$$L(t) = \begin{cases} L_{\mathrm{NS}} - L_1[1 - \exp(-t/\tau_1)] - L_2[1 - \exp(-t/\tau_2)] & t \geq 0 \\ L_{\mathrm{NS}} & t < 0. \end{cases} \tag{A2.3.4}$$

The time-dependent loss $L(t)$ includes two saturable contributions with strength L_1 and L_2, and a non-saturable term L_{NS}. The nonlinear response of equation (A2.3.4) can be measured using *pump-probe spectroscopy*. Pump-probe spectroscopy employs a second weaker pulse to monitor the depletion at variable delays after the strong pump pulse at $t = 0$. From a logarithmic plot of the measured response, one can easily retrieve the time constants τ_1 and τ_2. Figure A2.3.3(*e*) qualitatively shows the bitemporal behaviour. The induced transmission is strongest when all excited carriers are still in resonance with the pump that created them. Cooling processes rapidly broaden the initial hot distribution of electrons, with more and more electrons being transferred to the bottom of the conduction band. This is the first effect reducing the induced transmission signal. On longer time scales, the signal is further weakened by recombination back into the valence band until it will eventually vanish, which typically happens on the nanosecond time scale.

For applications of saturable absorbers in lasers, the time constants τ_1 and τ_2 are important device parameters. These time constants are to be compared with the pulse duration of the laser. Despite the fact that it is possible to generate pulses that are more than one order of magnitude shorter than the dominant response time of the absorber [24, 25], the usefulness of an absorber is greatly enhanced the faster its response. Clearly, a recombination time constant τ_2 of a few nanoseconds, i.e. approaching typical cavity roundtrip times, is problematic. Significant efforts went into *acceleration of the recombination process*. One method of doing so is the introduction of *additional mid-gap* states into the material as schematically shown in figure A2.3.3(*d*). These additional states can be generated by a variety of processes. The first observation that *defect-induced* additional states contribute to an accelerated relaxation of a saturable absorber was reported by Ippen *et al* [26]. Later several methods for introducing defects in a controlled way were explored, including low-temperature growth [27, 28], ion implantation [29], proton bombardment [30, 31], or impurity implantation [32]. Lifetimes have been pushed down to one to several picoseconds, which constitutes an improvement of about three orders of magnitude compared to low-defect materials.

A2.3.4.3 Design of saturable absorber mirrors

Apart from the time constants, the relative strength of the nonlinear absorption and its wavelength characteristics are further important design parameters of the SESAM. Both bulk absorption and *quantum wells* are used for generating saturable absorption in SESAMs. Quantum wells exhibit a relatively strong nonlinear response because of *quantum confinement* effects [21]. The number and position of intrinsic and excitonic states of a quantum well can be influenced both by the material and by the thickness of the quantum well. Note that saturable absorption extends to below the nominal band gap of a quantum well, as there are also strong *excitonic effects* in semiconductor heterostructures. Excitonic contributions tend to be faster but also weaker than contributions from band filling. The design of the quantum wells allows for some *tailoring of the wavelength response* and also of the strength of the saturable absorption. The latter can be further influenced by a multitude of design parameters, namely the number of quantum wells, their position relative to the nodes of the electric field, and also by the optical design of the SESAM [7]. If the quantum well is positioned at a maximum of the standing wave inside the laser cavity, its response is coupled much tighter to the electric field than if it were positioned at a node of the field (see figure A2.3.4).

Figure A2.3.4 shows two different aspects of the design of a SESAM device, its *bandgap structure* and its *refractive index structure*. The index structure is responsible for positioning of the nodes and the reflection properties [7]. The bandgap structure is engineered to provide suitable saturation properties,

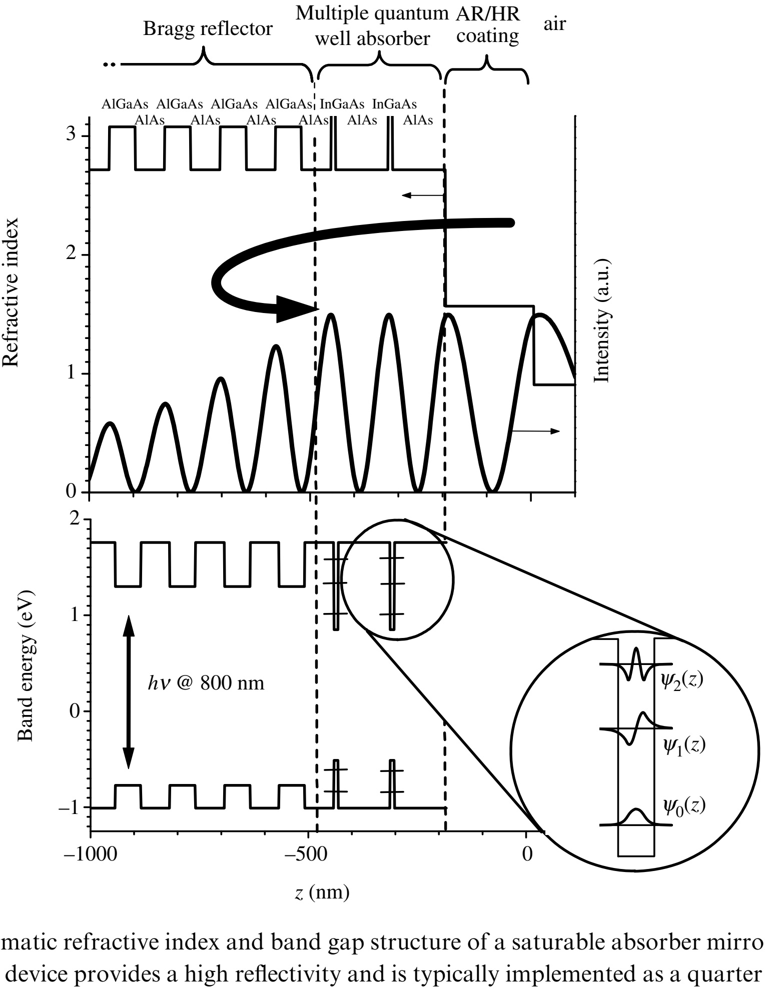

Bragg reflector
Multiple quantum well absorber
AR/HR coating
air
AlGaAs AlGaAs AlGaAs AlGaAs InGaAs InGaAs
AlAs AlAs AlAs AlAs AlAs AlAs
Refractive index
Intensity (a.u.)
3
2
1
0
Band energy (eV)
2
1
0
−1
$h\nu$ @ 800 nm
$\psi_2(z)$
$\psi_1(z)$
$\psi_0(z)$
−1000
−500
0
z (nm)

Figure A2.3.4. Schematic refractive index and band gap structure of a saturable absorber mirror. A Bragg reflector at the bottom of the device provides a high reflectivity and is typically implemented as a quarter wave layer stack of two different semiconductor materials. The index of refraction is alternating in this stack. Both materials should have a band gap larger than the photon energy in the device. Saturable absorption is produced by one or several quantum wells. In these quantum well regions, discrete states exist. Additionally excitonic effects may contribute to the absorption behaviour of the quantum wells. The quantum well position relative to the nodes of the field allows us to tailor the strength of the nonlinear response. The field strength at the quantum wells can be additionally manipulated by an external Fabry–Perot structure. For this purpose an additional mirror is coated on top of the device. This mirror can either be highly reflecting (HR, high-finesse case) or be an antireflection coating (AR).

as was explained above. As the SESAM has to also act as a mirror, the absorber structure is deposited on a highly reflecting mirror reflector. In the simplest case, a *Bragg mirror*, i.e. a combination of quarter-wave layers of alternating index of refraction, is used as outlined in figure A2.3.4. The bandgap of the materials used in the reflector is significantly higher than the photon energy, therefore negligible absorption takes place in the mirror structure. One of the disadvantages of the Bragg mirror approach is its limited bandwidth which is mainly caused by the weak index contrast of the available transparent semiconductor materials. Typically, a Bragg mirror consisting of AlGaAs and AlAs, as suggested in figure A2.3.4, only covers a bandwidth of about 100 nm at an 800-nm centre wavelength. Alternatively, metal mirrors have also been employed for providing more broadband reflectivity [33, 34]. This approach, however, is much more demanding in terms of manufacturing.

Apart from the Bragg reflector and the absorber, a second mirror structure can be added on top of the absorber towards the interface to air. The reflectivity of the top mirror can reach from more than 50% down to an antireflection coating. The case of high reflectivity results in a high finesse Fabry–Perot-like structure. Typically, high-finesse SESAMs are used in antiresonance (*antiresonant Fabry–Perot saturable absorber*, A-FPSA [35]). The antiresonance decreases the intensity of the light inside the absorber relative to the main cavity. For example, a 95% reflectivity top mirror increases the saturation fluence by about a factor 100 [7]. In more recent SESAM configurations, lower reflectivities of the top reflector were used and scaling of the nonlinearity was obtained by other means already discussed above [36]. The top mirror or antireflection structure can either be manufactured by semiconductor technologies or using dielectric materials such as TiO_2 or SiO_2. Very simple but efficient antireflection coatings can be manufactured by depositing a single half-wave layer of a material with index of refraction $n_{AR} = \sqrt{n_{top}}$, where n_{top} is the top layer refractive index.

Depending on the pulse duration being larger or smaller than the dominant time constant of the SESAM, one either observes a functional dependence of the reflectivity with pulse peak power or pulse energy, respectively. On a femtosecond time scale, the response of the SESAM generally scales with energy, because significant interband relaxation has not set in yet. In this case, one can write the power dependent reflectivity of a SESAM as

$$R(F) = R_0 + \Delta R(1 - \exp(-F/F_{sat})) \tag{A2.3.5}$$

where F is the fluence in Joule per unit area, F_{sat} is the saturation fluence, R_0 is the small-signal reflectivity, and ΔR is the modulation coefficient [7], which is connected to the coefficients L_1 and L_2 in equation (A2.3.4). The functional dependence of the reflectivity on the flux is shown in figure A2.3.3(*f*). At very high fluence $F > 10F_{sat}$, additional effects may contribute, which are treated below. For lower fluences, the saturable ΔR may range from L_2 to $L_1 + L_2$, depending on pulse duration. In the absence of non-saturable losses, $R_0 + \Delta R$ should add up to unity reflectivity. In reality, *non-saturable losses* $L_{NS} = 1 - R_0 - \Delta R$ are typically on the same order of magnitude as ΔR itself, i.e. a few percent. The onset of saturation is governed by the parameter F_{sat}. The saturation fluence is related to the *cross section* σ of the absorber by the relation

$$F_{sat} = \frac{h\nu}{N\sigma} \tag{A2.3.6}$$

i.e. the *saturation fluence* is reached when on average one photon with energy $h\nu$ impinges on the cross-sectional area σ. Inside a linear laser cavity, this may happen twice ($N = 2$). In a ring cavity or otherwise $N = 1$. The saturation fluence is a decisive engineering parameter of the SESAM, and should be characterized carefully prior to use of a SESAM inside a laser. The saturation fluence of the absorber has to be carefully balanced with the saturation fluence of the laser gain material to avoid spiking of the laser. The gain saturation fluence is related to the gain cross section as defined in equation (A2.3.6) [37].

For clean mode-locked pulse trains, it is of paramount importance to scale spot sizes in the gain medium and on the absorber accordingly. Guidelines for a suitable layout of a laser cavity are given in [38].

A2.3.4.4 Two-photon absorption in saturable absorber mirrors

For extremely high peak intensities, the nonlinear response of a saturable absorber may be affected by *TPA* giving rise to an additional term in equation (A2.3.4) [39–41]. As TPA is a quasi-instantaneous nonlinear process (see figure A2.3.1), this mechanism always scales with peak intensity $I(t)$, regardless of the pulse duration regime.

$$\Delta R_{\mathrm{TPA}} = -\beta_2 I. \tag{A2.3.7}$$

As TPA creates additional absorption for high enough peak powers, it defeats the effect of saturable absorption at extremely high power levels, as shown in figure A2.3.3(*f*). At first sight, TPA may appear as a limitation of SESAMs. However, this effect can be beneficial, causing a stabilization of the pulse duration and preventing pulse break up, stabilizing the laser pulse duration and the average power of the laser. This effect can also reduce the spiking tendency of the laser [41] and help to provide clean mode-locked pulse trains with stable pulse energy.

A2.3.4.5 Devices closely related to saturable absorber mirrors

The SESAM has found widespread use as intracavity saturable absorbers for generating short optical pulse [7, 23]. Several similar devices have been proposed and demonstrated. The SESAM's main application is *self-amplitude modulation*. In a slightly modified geometry, saturable absorption can also be used to modulate one laser beam with the aid of another one [42]. Here the control light is fed into a waveguide and bleaches out an arrangement of several quantum wells. This saturable absorption modulates a light beam that crosses the quantum wells at normal incidence to the waveguide direction.

Other nonlinear optical switches have been proposed, which do not directly rely on the absorptive properties of quantum wells or semiconductor bulk materials. These devices embed the saturable absorber in between a high-finesse Fabry–Perot structure. Other than with the A-FPSA approach previously mentioned, the Fabry–Perot is used at resonance and relies on refractive index changes accompanying the creation of free carriers in the conduction band of the semiconductor. While SESAMs can be understood as amplitude modulators, this type of *Fabry–Perot modulator* (FPM) relies on *phase modulation*. For inducing mode-locking inside a laser resonator, the SPM has to be converted back into an amplitude modulation, compare 'Nonlinear interferometers as switches' section. This is accomplished by the Fabry–Perot structure. The index change causes a tuning of the Fabry–Perot resonances. If the length is chosen such that the additional index change increases the overall reflectivity of the device, such a device can also be used as an optical switch for mode-locking [43]. Moreover, one can further exploit the phase modulation aspect of such Fabry–Perot structures for synchronizing lasers as shown in [44]. Using a laser with a photon energy above the band gap of the spacer material in the FPM, one can modulate index of refraction for light below gap without inducing any losses. This index modulation allows control of the cavity roundtrip time and passive locking of cavity repetition rates.

A2.3.4.6 Inverse saturable absorption—optical limiting

For many applications, an effect with exactly the reverse behaviour of saturable absorption is desirable. One example could be protection of a sensitive photo-detector from intense light pulses that might otherwise destroy the detector. Such a device requires a transmission that decreases with increasing light intensities, a characteristic that is referred to as *optical limiting*. The previous discussion already addressed *two-photon absorption* in semiconductor materials as one possible mechanism showing exactly the right sign of saturable absorption needed for an optical limiter ([39–41], see equation (A2.3.7) and figure A2.3.3(*f*)). In fact, semiconductor materials such as GaAs have been suggested early on for this purpose [45, 46]. Additionally *self-defocusing* may contribute to optical limiting in these devices, an effect that will be explained in the following section. Rather than using semiconductors, one can again also use dyes for the same purpose [44–46]. One of the major mechanisms for reverse saturable absorption in complex organic molecules is the generation of excited electronic states with an absorption cross section larger than the ground state absorption. *Excited state absorption* has been reported for several classes of organic chromophores [47]. Apart from sequential single-photon absorption, TPA may also contribute to optical limiting in dye molecules [48, 49]. Different chemical compounds, such as fullerenes, organometallics, or carbon black suspensions have been discussed for optical limiting devices. Optically induced scattering, e.g. by localized melting of the limiter material, has also been suggested as a mechanism for optical limiting [50]. However, in spite of the variety of nonlinearities, materials, and device configurations that have been used to implement passive optical limiters, no single device or combination of devices has yet been identified that will protect any given sensor from all potential optical threats.

A2.3.5 The all-optical Kerr effect and nonlinear refraction

A2.3.5.1 Kerr-based switches and Kerr gates

From the point of view of optoelectronics, saturable absorption is probably the most interesting process among the variety of nonlinear optical processes, as it allows us to build an *all-optical switch*. Depleting carriers by means of a pump pulse, one can influence the transmission properties of a saturable absorber and thereby control the energy of a second probe pulse. In some way, this device may be thought as an all-optical transistor or *light valve*, despite the fact that one needs rather large light intensities to control relatively small ones. Apart from acting as a light valve, a technologically important aspect of an all-optical switch is that it forms the product of optical waveforms. Multiplying an unknown waveform with a known one is at the heart of optical sampling techniques and allows characterization of unknown optical pulse shapes with the aid of known ones. Unfortunately, with SESAMs one can do neither of these arbitrarily fast, as saturable absorption is a rather slow process with picosecond relaxation times even when midstate traps are used to accelerate this process. This calls for methods that benefit from *quasi-instantaneous nonlinearities* (cf 'Quasi-instantaneous nonlinear optical processes' section).

It is rather clear that the temporal response of carrier-related effects cannot be much further accelerated without sacrificing modulation depth. Instead, many approaches have been pursued to employ bound-electronic optical processes to build an all-optical switch. These approaches rely nearly exclusively on *nonlinear refraction*, i.e. the change of the index of refraction with instantaneous intensity. First and foremost, nonlinear refraction is a phase effect that requires *translation into an amplitude effect* to be effective as a nonlinear optical switch. In the following, we will first review the physics of nonlinear refraction and point out that it is inseparably connected to TPA and other nonlinear optical effects. In fact, this connection between nonlinear refraction and TPA poses a stringent limitation of all-optical valves. We will then give an overview of several architectures of all-optical switches based on nonlinear refraction. These switches have found widespread use in rather distinct applications of optics ranging from telecommunications to measurement techniques for short pulses.

A2.3.5.2 The physics of nonlinear refraction

The propagation (phase) velocity of light v_φ in a dielectric medium is governed by the refractive index n according to $v_\varphi = c/n$. The refractive index n is a function of wavelength and can be computed from the complete knowledge of absorbing resonances in the ultraviolet and infrared spectral region using *Kramers–Kronig relationship* [51]. The slightest changes, e.g. of the lattice constant or temperature of a solid-state material, will automatically affect the refractive index when they modify the spectral position or relative strength of the resonances. In dielectric media this will cause the index of refraction change according to

$$n(I) = n_0 + n_2 I. \tag{A2.3.8}$$

This effect of an intensity-dependent refractive index is called the *all-optical Kerr effect* [1, 3, 4]. The coefficient n_2 is called the *nonlinear index of refraction* and is related to the real part of the susceptibility tensor $\chi^{(3)}$ via

$$n_2 = \frac{3}{8n} Re\chi^{(3)} \tag{A2.3.9}$$

where it was assumed that the light is linearly polarized such that only one component of the third-rank tensor $\chi^{(3)}$ contributes. Often, $\chi^{(3)}$ is quoted using electrostatic units (esu). One can convert to the more useful W cm^{-2} employing the relation

$$n_2(\mathrm{cm}^2\,\mathrm{W}^{-1}) = \frac{12\pi^2}{n_0^2 c} Re\chi^{(3)}(\mathrm{esu}). \tag{A2.3.10}$$

Typical values for n_2 of optical glasses lie in the range of a few $10^{-16}\,\mathrm{cm}^2\,\mathrm{W}^{-1}$ [4, 52, 53]. The main effect in a dielectric medium far away from the bandgap is electronic polarization, which has a response time of 1 fs or faster. In other materials, additional effects such as molecular orientation or atomic resonances can create much stronger nonlinear refraction, at the expense of a greatly slowed response and a much narrower bandwidth. In the following, we will restrict ourselves on the Kerr effect in dielectrics and semiconductors and focus on nonlinear refraction induced by electronic polarization.

A power-dependent index of refraction means that the phase velocity of the light has become a function of intensity, with high peak power pulses travelling slower than a low peak power beam with identical cross section would do, see equation (A.2.3.8). This self-action of the light is called SPM

$$\Delta\varphi(I) = \frac{2\pi}{\lambda} n_2 I L \tag{A2.3.11}$$

where λ is the length of the medium and L the wavelength of the light. The importance of nonlinear refraction can be estimated from inverting equation (A2.3.11) to find the critical intensity that is necessary to create a macroscopic 2π phase shift

$$I_{\mathrm{crit}} = \frac{\lambda}{n_2 L}. \tag{A2.3.12}$$

Plugging in numbers for a 1 cm long piece of glass, one finds a critical intensity level on the order of $10^{11}\,\mathrm{W\,cm}^{-2}$. This makes it immediately clear that SPM only becomes important when laser pulses are either extremely tightly focused or when a very long interaction length can be used as in optical fibres. In fact, SPM is extensively used in *nonlinear fibre optics* [4].

A2.3.5.3 Connection of nonlinear refraction and TPA

As nonlinear refraction is the most important effect for building quasi-instantaneous optical switches, a somewhat deeper insight into the underlying physics is important to gain an understanding on some of the limitations. The Kerr effect can be generated by a variety of different mechanisms, including molecular orientation, saturated absorption, and electrostriction [1]. In a wider sense, also thermal effects create refractive index changes. With response times of picosecond to milliseconds, application of these effects is much less interesting than use of the Kerr effect induced by electronic polarization changes. For all practical applications this response can be considered instantaneous. Quite generally, there appears to be wide-ranging inverse connection between the response time of the effect and its relative strength. Investigations of the electronic Kerr effect therefore were motivated by finding a material that combines quasi-instantaneous response with maximum nonlinear susceptibility.

In semiconductors, an *inverse Kerr effect* has been observed, i.e. high intensities travel at higher velocity than low intensities, other than in dielectric materials. This sign change can normally be explained by the *generation of free carriers*, which is a strong indication for the onset of multiphoton absorption in the material. The magnitude of the nonlinear response shows a dramatic increase close to the band gap of an optical material [54]. This renders semiconductors very interesting for building all-optical switches. Figure A2.3.5 shows the fundamental behaviour of nonlinear refraction and TPA *vs*. photon energy. This curve is based on the theoretical model of Sheik-Bahae *et al* for nonlinear refraction. The computation is based on the *Kramers–Kronig relationship*, which allows us to calculate refraction changes directly from absorption changes [51, 55]. The main effect contributing to the nonlinear absorption is *two-photon absorption*, which sets in above half the band gap of the material.

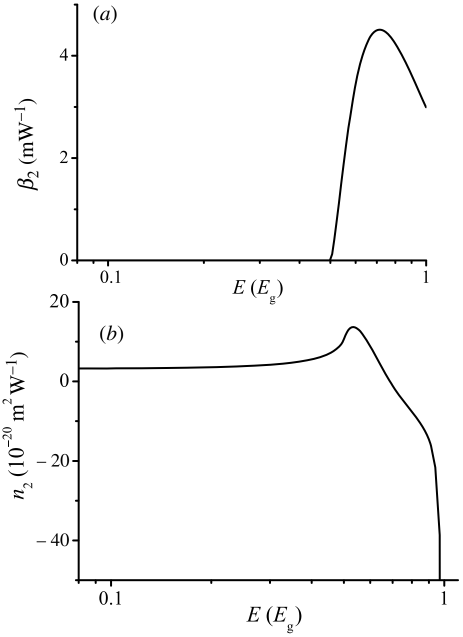


Figure A2.3.5. Behaviour of two-photon absorption (*a*) and nonlinear refraction (*b*) as a function of photon energy. These calculations are based on material parameters of silica. The photon energy is scaled in units of the band gap of the material. Nonlinear refraction is rather constant at low energies, exhibits a peak above half the band gap and then turns negative close to the band gap.

Other effects, such as the *Stark effect* and *Raman* contributions, also play a role and have been accounted for in figure A2.3.5. It is important to understand that TPA poses a severe limitation on using the Kerr effect for a switching application. Excessive TPA will destroy the switch. From the behaviour of nonlinear refraction alone, one would tend to go as close to the band gap as possible to create the maximum nonlinear response. Unfortunately, such a device would also suffer from strong possible TPA. For switching applications, it has been found that an optimum between both effects is reached when the photon energy is about 70% of the band gap energy [55].

A2.3.5.4 The Kerr lens as a nonlinear optical switch

The earliest switching method that was based on the Kerr effect is the so-called Kerr gate [56], which is illustrated in figure A2.3.6. In a Kerr gate, a probe pulse is sent through a Kerr medium, in which its path crosses the path of a strong pump pulse. Crossed polarizers prevent detection of the probe light in the absence of an overlap with the pump. Only if the pump induces a *polarization rotation* inside the Kerr medium, is some portion of the probe transmitted at the polarizer. This Kerr gate is instantaneous, provided that the Kerr nonlinearity is sufficiently fast. In the more recent literature, interferometer based switches are also referred to as optical gates (see e.g. [57]). We will treat these methods of translating

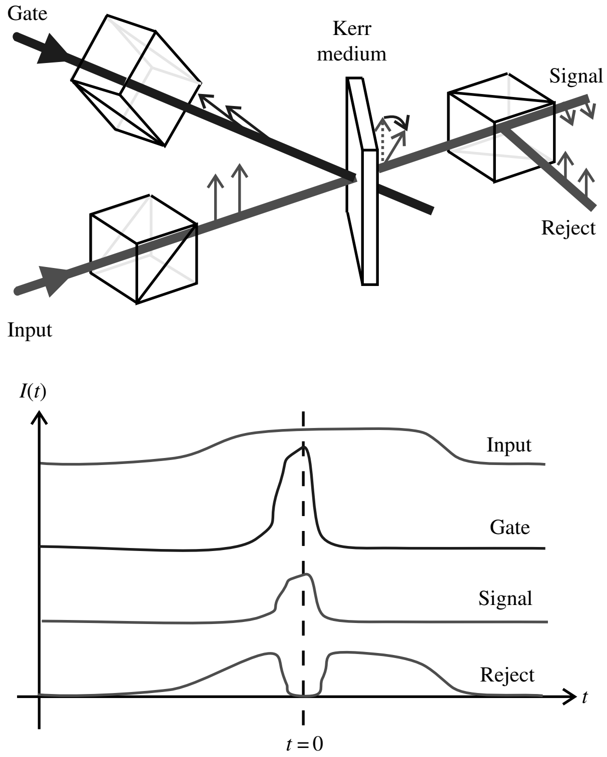


Figure A2.3.6. Kerr gate based on polarization rotation. The gating pulse and the input pulse are crossed inside a Kerr medium. Both pulses are linearly polarized, with the polarization of the gating pulse rotated 45° relative to the input pulse. The Kerr effect causes a polarization rotation of the input pulse. A second crossed polarizer in the input beam line allows separation of the gated pulse from the rejected part of the input pulse. An example for the temporal behaviour of the gating pulse, the input pulse, and the two output pulses are shown below for comparison.

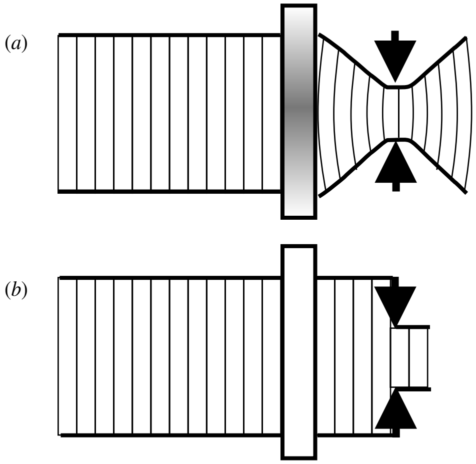


Figure A2.3.7. Kerr gate based on Kerr lensing. Schematically shown are propagating wavefronts of a high-intensity and a low-intensity beam, (*a*) and (*b*), respectively. For sufficiently high intensities, an transverse index profile is induced by the pulse as indicated by grey shades. The stronger the induced change of the refractive index, the stronger the corresponding part of the beam front is retarded. This causes an additional beam front curvature, i.e. the Kerr effect transversally acts like a lens. A subsequent aperture can be used to convert the lensing effect into an effective amplitude modulation. The combination of the Kerr lens and an aperture acts similar to the gate of figure A2.3.6, but is not background-free.

an instantaneous phase nonlinearity into an amplitude nonlinearity below. Instantaneous optical gates can be built from Kerr induced polarization changes, Kerr lenses, or nonlinear interferometers.

A further Kerr-related effect that can be used for building a switch is *self-focusing*. This method is illustrated in figure A2.3.7 [3, 58, 59]. When the central most intense part of a beam experiences a phase delay according to equation (A.2.3.8), this is equivalent to the focusing action of a convex lens. In a convex lens the phase shift is generated by the larger amount of glass in the beam path of the central rays. Other than in an ideal conventional lens, however, the wavefront change induced by the nonlinear Kerr lens is not exactly spherical, but proportional to the spatial beam profile. For example, if the beam profile is Gaussian, only the central part of the Kerr lens can be approximated by a conventional lens. A nonlinear *gradient-index duct* [37, 59, 60] is normally a better approximation for the Kerr lens.

As the Kerr lens only focuses the most intense parts of a short pulse, one can use it to build an ultrafast optical switch, which is also illustrated in figure A2.3.7. If an aperture is suitably placed in the focus of the Kerr lens, it will introduce losses for cw laser light of low intensity, whereas high intensity light can pass through as it sees the focusing action of the Kerr lens. This kind of self-switching is used in lasers to discriminate short pulses and create losses for any kind of low intensity background. Placed in a cavity, the Kerr lens switch strips off a pulse pedestal, recleaning the pulse on every passage. This is used for generating some of the shortest pulses ever generated from a laser with a method called *Kerr-lens mode-locking* [59–62].

A2.3.5.5 Nonlinear interferometers as switches

Another method to build a switch is based on SPM (compare equation (A2.3.11)). Here the Kerr effect is translated into an effective SAM in a *nonlinear interferometer*. The basic idea is illustrated in figure A2.3.8 with a Michelson interferometer. If one introduces a Kerr medium in only one arm of the interferometer, high light intensities will experience a phase shift compared to low intensities. Provided

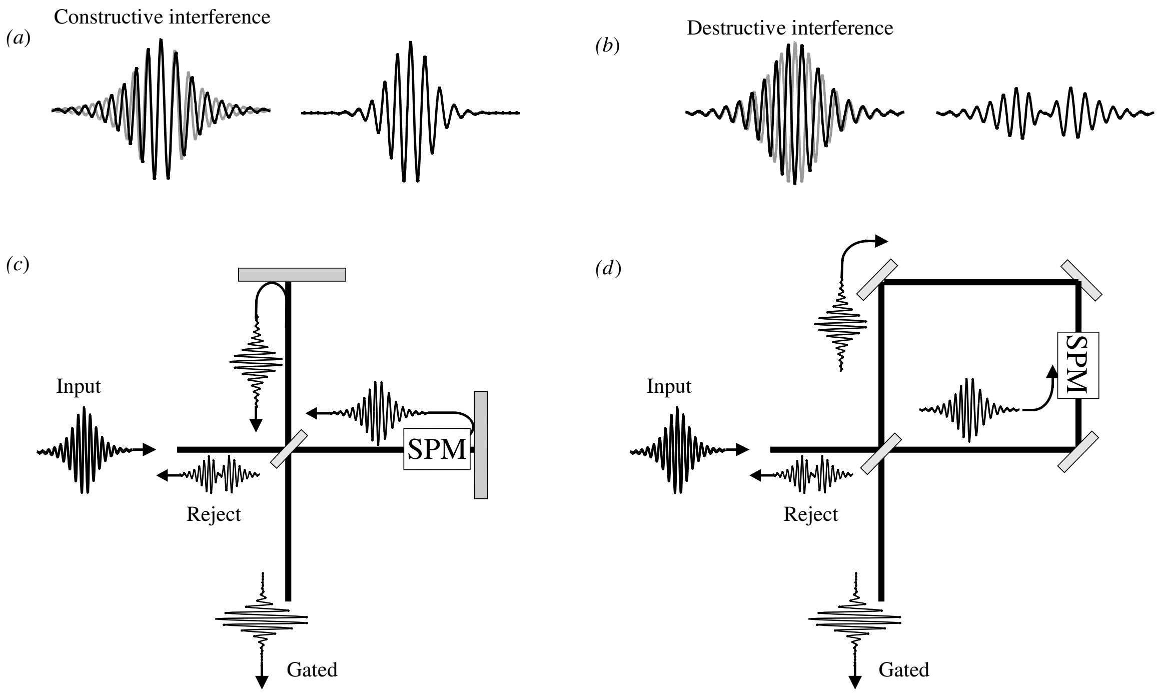


Figure A2.3.8. Kerr gate based on nonlinear interferometers. (*a*) and (*b*) show interference of an unmodulated pulse (grey line) with a self-phase modulated pulse (black line). If the bias phase at pulse centre is chosen for constructive interference as in (*a*), destructive interference in the wings compresses the resulting pulse. The opposite case is shown in (*b*), resulting in a lengthened pulse. (*c*) shows a Michelson configuration, where cases (*a*) and (*b*) appear in the two different output ports of the interferometer. A Sagnac configuration (*d*) works in a similar way, at the added advantage that no stabilization of the bias phase is required.

the bias phase and the nonlinear phase are properly set, low intensities will be back-reflected into the input port whereas high intensities go into the opposite port. Sending pulses in such a nonlinear Michelson interferometer, only the central and most intense part of the pulse will go into the output port. The weaker parts of the pulse will be redirected into the input port, effectively leading to a compression of the input pulse. A Michelson interferometer is not the ideal configuration for such a nonlinear switch, because it will require active means to stabilize the bias phase of the interferometer. Mach–Zehnder interferometers can be built in a micro-optical configuration, dramatically reducing problems of phase drift [57]. A *Sagnac* configuration, as has first been suggested by Blow and Doran [63], is typically the method of choice. In the *nonlinear optical loop mirror* (NOLM) the two interferometer arms consist of different direction of propagation in a fibre loop (see figure A2.3.8). An imbalance of the Kerr effect between the two directions of propagation is simply induced by asymmetric beam splitting. Otherwise the function is identical to the previously discussed case. The high-intensity portion of the pulse is switched into one port of the interferometer, the pedestal goes into the other port, effectively leading to pulse compression. Such Sagnac interferometers have been used for pulse shape cleaning and also to induce mode-locking.

As the impression may arise that the exact type of interferometer does not matter for the principle of operation, it needs to be pointed out that the interferometer types discussed so far only exploit dual-beam interference. Using a Fabry–Perot interferometer instead, interference of many pulses may create complex behaviour. A nonlinear Fabry–Perot switch has been extensively discussed as an all-optical switch as it exhibits bistability and hysteresis, a behaviour analogous to what a *Schmitt trigger* does in electronics. Using nonlinear systems with large feedback is interesting for applications as it allows us to switch large light powers with relatively small ones. However, this method has not found very widespread use because of several disadvantages. The nonlinear Fabry–Perot has extremely rich nonlinear dynamics, including *period-doubling* and *chaotic behaviour* [64, 65]. Generally, this type of behaviour becomes more pronounced the stronger the feedback inside the system is. Another disadvantage of all nonlinear interferometers is the delay of the interferometer arms. In the case of two-pulse interference the actual switching action happens after single propagation through the interferometer. In the case of the nonlinear Fabry–Perot, the switching time amounts to many roundtrips through the interferometer. However, despite these problems, nonlinear Fabry–Perot interferometers have been successfully incorporated into lasers to initiate and sustain mode-locking [66–69]. Compared to saturable-absorber mode-locking or Kerr-lens mode-locking, these lasers require a stabilization of the interferometer phase.

A2.3.5.6 Applications of nonlinear optical switches

One important application of a nonlinear optical switch has already been addressed in the previous section—mode-locking. For generation of short pulses the nonlinear optical switch is inserted into the laser cavity. Its main purpose is the *suppression of low light levels* and a preference for high light levels. Such a preferred transmission of high peak powers leads to a temporal focusing of the entire energy contents of the cavity into one short pulse. Continuous operation of the cavity is discouraged by the higher losses induced by the nonlinear switch. As the switch reduces the power in the pulse's pedestals compared to its peak power, one may also think of the switch *recompressing the pulse* every round-trip [68]. The nonlinear switch initializes the mode-locking process from noise and stabilizes it in the steady-state regime. Additional processes such as dispersion and spectral broadening contribute during the steady-state operation and determine the pulse duration of the mode-locked laser.

A second application of a nonlinear switch is *pulse characterization*. One can use the geometry of the Kerr gate, e.g. to sample one unknown pulse with the aid of another one, simply by varying the delay between pump and probe in the Kerr gate. This is known as *cross-correlation*. Moreover, one can also try to sample one pulse by itself, i.e. perform an *autocorrelation*. Pulse retrieval from autocorrelation is ambiguous and impossible without further knowledge on the pulse shape. Advanced methods to measure and characterize pulse shapes are discussed in chapter B7.

Telecommunication also strongly relies on optical switches. For many applications, electro-optic switches are still fast enough for data processing. However, there is a strong tendency to substitute electro-optical methods by all-optical variants, in particular when approaching data rates in the tens of Gbit s^{-1}. One potential application lies in the area of *time-domain multiplexing and demultiplexing* of optical data streams. One can use fibre-based switches similar to those previously discussed to selectively extract, e.g. every fourth or sixteenth pulse from a stream of pulses. The extracted pulse train then has a much smaller data rate than the original one and can be processed with slower electro-optic components. An optical switch based on a polarization discriminating Mach–Zehnder interferometer has been demonstrated with up to 1500 GB s^{-1} data rates and 200-fs pulse duration [70]. Many different switch architectures have been proposed for demultiplexing applications [42, 71–75]. Another application of all-optical switches is optical memories. It has been demonstrated that pulse sequences can be *stored* inside fibre ring resonators. A pulse sequence is launched into the ring and later extracted by an optical switch [76, 77]. For launching or extracting an entire sequence at once, no particular fast switch is needed. With an all-optical switch, however, individual pulses can be launched or extracted from the optical ring buffer, which greatly enhances the utility of such optical memories.

A2.3.6 Summary

Optical nonlinearities are important building blocks in optoelectronic devices, which have a variety of different applications as optical switches. Two fundamentally different basic concepts have evolved. One of them is the *saturable absorber*, which is based on saturation of an electronic transition. Given the very strong interaction of quantum wells and light, extremely compact all-optical switches can be built using modern semiconductor manufacturing technology. These devices have found widespread application for mode locking of lasers. Their only limitation is the limited switching speed, which cannot easily be pushed below 1 ps. A second concept employs *quasi-instantaneous nonlinearities*, namely the *Kerr effect*, to overcome speed limitations of the device. As this nonlinearity is a phase-nonlinearity, some means of phase to amplitude conversion has to be incorporated into the device to make it an effective saturable absorber with practically unlimited switching speed. This translation can be accomplished either by an interferometer, by polarization switching, or by using the Kerr lens mechanism. As the Kerr effect is a third-order nonlinearity, it is typically much weaker than saturable absorption of a quantum well. Often, sufficient nonlinearity can only be accumulated by propagation through a long piece of fibre. Therefore, many concepts for using the Kerr nonlinearity as a switch involve relatively long delays compared to the more directly switching semiconductor-based saturable absorbers. It strongly depends on the application, which of the two fundamental concepts is to be preferred.

References

[1] Boyd R W 1992 *Nonlinear Optics* (San Diego, CA: Academic Press)
[2] Dmitriev V G, Gurzadyan G G and Nikogosyan D N 1999 *Handbook of Nonlinear Optical Crystals* 3rd edn Springer Series in Optical Sciences **64** (Berlin: Springer)
[3] Yariv A 1989 *Quantum Electronics* 3rd edn (New York: Wiley)
[4] Agrawal G P 2001 *Nonlinear Fiber Optics* 3rd edn (San Diego, CA: Academic Press)
[5] Keller U, 't Hooft G W, Knox W H and Cunningham J E 1991 Femtosecond pulses from a continuously self-starting mode-locked Ti:sapphire laser *Opt. Lett.* **16** 1022–1024
[6] Tsuda S, Knox W H, De Souza E A, Jan W Y and Cunningham J E 1995 Low-loss intracavity AlAs/AlGaAs saturable Bragg reflector for femtosecond mode-locking in solid-state lasers *Opt. Lett.* **20** 1406–1408
[7] Keller U, Weingarten K J, Kärtner F X, Kopf D, Braun B, Jung I D, Fluck R, Hönninger C, Matuschek N and Aus der Au J 1996 Semiconductor saturable absorber mirrors (SESAM's) for femtosecond to nanosecond pulse generation in solid-state lasers *IEEE J. Sel. Top. Quantum Electron.* **2** 435–453
[8] Hillman L W and F J Duarte (eds) 1990 *Dye Laser Principles: With Applications* (Boston, MA: Academic Press)
[9] Duarte F J (ed) 1992 *Selected Papers on Dye Lasers* SPIE Milestone Series **45** (Boca Raton, FL: SPIE Press)
[10] Bilinsky I P, Fujimoto J G, Walpole J N and Missaggia L J 1999 InAs-doped silica films for saturable absorber applications *Appl. Phys. Lett.* **74** 2411–2413
[11] Wu E, Chen H, Sun Z and Zeng H 2003 Broadband saturable absorber with cobalt-doped tellurite glasses *Opt. Lett.* **28** 1692–1694
[12] Okhrimchuk A G and Shestakov A V 1994 Performance of YAG: Cr^{4+} laser crystal *Opt. Mater.* **3** 1–13
[13] Shimony Y, Burshtein Z and Kalisky Y 1995 Cr^{4+}:YAG as passive Q-switch and Brewster plate in a pulsed Nd:YAG laser *IEEE J. Quantum Electron.* **31** 1738–1741
[14] Eilers H, Hoffman K R, Dennis W M, Jacobsen S M and Yen W M 1992 Saturation of 1.064 μm absorption in Cr,Ca:$Y_3Al_5O_{12}$ crystals *Appl. Phys. Lett.* **61** 2958–2960
[15] Griebner U and Koch R 1995 Passively Q-switched Nd:glass fibre-bundle laser *Electron. Lett.* **21** 205–206
[16] Gibson A F, Kimmit M F and Norris B 1974 Generation of bandwidth-limited pulses from a TEA CO_2 laser using p-type germanium *Appl. Phys. Lett.* **24** 306–307
[17] Boyd G D, Bowers J E, Soccolich C E, Miller D A B, Chemla D S, Chirovsky L M F, Gossard A C and English J H 1989 5.5 GHz multiple quantum well reflection modulator *Electron. Lett.* **25** 558–560
[18] Chmielowski M and Langer D W 1989 Quantum well vertical light modulator *IEEE Photon. Technol. Lett.* **1** 77–79
[19] Whitehead M and Parry G 1989 High-contrast reflection modulation at normal incidence in asymmetric multiple quantum well Fabry–Perot structure *Electron. Lett.* **25** 566–568
[20] Klingshirn C F 1995 *Semiconductor Optics* (Berlin: Springer)
[21] Weisbuch C and Vinter B 1991 *Quantum Semiconductor Structures—Fundamentals and Applications* (San Diego, CA: Academic Press)

[22] Knox W H, Hirlimann C, Miller D A B, Shah J, Chemla D S and Shank C V 1986 Femtosecond excitation of nonthermal carrier populations in GaAs quatum wells *Phys. Rev. Lett.* **56** 1191–1193

[23] Keller U 2003 Recent developments in compact ultrafast lasers *Nature* **424** 831–838

[24] Haus H A 1975 Theory of mode-locking with a slow saturable absorber *IEEE J. Quantum Electron.* **11** 736–746

[25] Kärtner F X, Jung I D and Keller U 1996 Soliton mode-locking with saturable absorbers *IEEE J. Sel. Top. Quant.* **2** 540–556

[26] Ippen E P, Eichenberger D J and Dixon R W 1980 Picosecond pulse generation by passive modelocking of diode lasers *Appl. Phys. Lett.* **37** 267–269

[27] Gupta S, Whitaker J F and Mourou G A 1992 Ultrafast carrier dynamics in II–V-semiconductors grown by molecular beam epitaxy at very low substrate temperatures *IEEE J. Quantum Electron.* **28** 2464–2472

[28] Siegner U, Fluck R, Zhang G and Keller U 1996 Ultrafast high-intensity nonlinear absorption dynamics in low-temperature grown gallium arsenide *Appl. Phys. Lett.* **69** 2566–2568

[29] Tan H H, Jagadish C, Lederer M J, Luther-Davies B, Zou J, Cockayne D J H, Haiml M, Siegner U and Keller U 1999 Role of implantation-induced defects on the response time of semiconductor saturable absorbers *Appl. Phys. Lett.* **75** 1437–1439

[30] van der Ziel J P, Tsang W T, Logan R A, Mikulyak R M and Augustyniak W M 1981 Subpicosecond pulses from passively modelocked GaAs buried optical guide semiconductor lasers *Appl. Phys. Lett.* **39** 525–527

[31] Gopinath J T, Thoen E R, Koontz E M, Grein M E, Kolodziejski L A, Ippen E P and Donnelly J P 2001 Recovery dynamics in proton-bombarded semiconductor saturable absorber mirrors *Appl. Phys. Lett.* **78** 3409–3411

[32] Delpon E L, Oudar J L, Bouché N, Raj R, Shen A, Stelmakh N and Lourtioz J M 1998 Ultrafast excitonic saturable absorption in ion-implanted InGaAs/InAlAs multiple quantum wells *Appl. Phys. Lett.* **72** 759

[33] Fluck R, Jung I D, Zhang G, Kärtner F X and Keller U 1996 Broadband saturable absorber for 10-fs pulse generation *Opt. Lett.* **21** 743–745

[34] Zhang Z, Nakagawa T, Torizuka K, Sugaya T and Kobayashi K 2000 Gold-reflector-based semiconductor saturable absorber mirror for femtosecond mode-locked Cr4+: YAG lasers *Appl. Phys.* B **70** S59–S62

[35] Keller U, Miller D A B, Boyd G D, Chiu T H, Ferguson J F and Asom M T 1992 Solid-state low-loss intracavity saturable absorber for Nd:YLF lasers: an antiresonant semiconductor Fabry–Perot saturable absorber *Opt. Lett.* **17** 505–507

[36] Brovelli L R, Keller U and Chiu T H 1995 Design and operation of antiresonant Fabry–Perot saturable semiconductor absorbers for mode-locked solid-state lasers *J. Opt. Soc. Am.* B **12** 311–322

[37] Siegman A E 1986 *Lasers* (New York: University Science Books)

[38] Hönninger C, Paschotta R, Morier-Genoud F, Moser M and Keller U 1999 Q-switching stability limits of cw passive modelocking *J. Opt. Soc. Am.* B **16** 46–56

[39] Thoen E R, Koontz E M, Joschko M, Langlois P, Schibli T R, Kärtner F X, Ippen E P and Kolodziejski L A 1999 Two-photon absorption in semiconductor saturable absorber mirrors *Appl. Phys. Lett.* **74** 3927–3929

[40] Langlois P, Joschko M, Thoen E R, Koontz E M, Kärtner F X, Ippen E P and Kolodziejski L A 1999 High fluence ultrafast dynamics of semiconductor saturable absorber mirrors *Appl. Phys. Lett.* **75** 3841–3843

[41] Schibli T R, Thoen E R and Kärtner F X 2000 Suppression of Q-switched mode locking and break-up into multiple pulses by inverse saturable absorption *Appl. Phys.* B **70** S41–S49

[42] Guina M D, Vainionpaa A, Orsila L, Harkonen A, Lyytikainen J, Gomes L A and Okhotnikov O G 2003 Saturable absorber intensity modulator *IEEE J. Quantum Electron.* **39** 1143–1149

[43] Seitz W, Ell R, Morgner U, Schibli T R, Kärtner F X, Lederer M J and Braun B 2002 All-optical active mode locking with a nonlinear semiconductor modulator *Opt. Lett.* **27** 2209–2211

[44] Seitz W, Schibli T R, Morgner U, Kärtner F X, Lange C H, Richter W and Braun B 2002 Passive synchronization of two independent laser oscillators with a Fabry–Perot modulator *Opt. Lett.* **27** 454–456

[45] Boggess T F, Moss S C, Boyd I W and Van Stryland E W 1985 Optical limiting in GaAs *IEEE J. Quantum Electron.* **21** 488–494

[46] van Stryland E W, van her Zeele H, Woodall M A, Soileau M J, Smirl A L, Guha S and Boggess T F 1985 2-photon-absorption, nonlinear refraction, and optical limiting in semiconductors *Opt. Eng.* **24** 613–623

[47] Perry J W, Mansour K, Marder S R, Perry K J, Alvarez D and Choong I 1994 Enhanced reverse saturable absorption and optical limiting in heavy-atom-substituted phtalocyanines *Opt. Lett.* **19** 625–627

[48] He G S, Xu G C, Prasad P N, Reinhardt B A, Bhatt J C and Dillard A G 1995 2-photon absorption and optical-limiting properties of novel organic-compounds *Opt. Lett.* **20** 435–437

[49] Ehrlich J E, Wu X L, Lee I Y S, Hu Z Y, Rockel H, Marder S R and Perry J W 1997 Two-photon absorption and broadband optical limiting with bis-donor stilbenes *Opt. Lett.* **22** 1843–1845

[50] Tutt L W and Boggess T F 1993 A review of optical limiting mechanisms and devices using organics, fullerenes, semiconductors and other materials *Prog. Quantum Electron.* **17** 299–338

[51] Nussenzweig H M 1972 *Causality and Dispersion Relations* (New York: Academic Press)

[52] Stolen R H and Ashkin A 1973 Optical Kerr effect in glass waveguide *Appl. Phys. Lett.* **22** 294–296

[53] Sheik-Bahae M, Said A A, Wei T H, Hagan D J and Van Stryland E W 1990 Sensitive measurement of optical nonlinearities using a single beam *IEEE J. Quantum Electron.* **26** 760–769

[54] De Salvo R, Said A A, Hagan D J, Van Stryland E W and Sheik-Bahae M 1996 Infrared to ultraviolet measurements of two-photon absorption and n_2 in wide bandgap solids *IEEE J. Quantum Electron.* **32** 1324–1333

[55] Sheik-Bahae M, Hutchings D C, Hagan D J and Van Stryland E W 1991 Dispersion of bound electronic nonlinear refraction in solids *IEEE J. Quantum Electron.* **27** 1296–1309

[56] Duguay M A and Hansen J W 1969 An ultrafast light gate *Appl. Phys. Lett.* **15** 192–194

[57] Lattes A, Haus H A, Leonberger F J and Ippen E P 1983 An ultrafast all-optical gate *IEEE J. Quantum Electron.* **19** 1718–1723

[58] Kelley P L 1965 Self-focusing of optical beams *Phys. Rev. Lett.* **15** 1005

[59] Salin F, Squier J and Piché M 1991 Mode-locking of Ti:Al_2O_3 lasers and self-focusing—a Gaussian approximation *Opt. Lett.* **16** 1674–1676

[60] Magni V, Cerullo G, De Silvestri S and Monguzzi A 1995 Astigmatism in Gaussian-beam self-focusing and in resonators for Kerr-lens mode-locking *J. Opt. Soc. Am.* B **12** 476–485

[61] Spence D E, Kean P N and Sibbett W 1991 60-fsec pulse generation from a self-mode-locked Ti:sapphire laser *Opt. Lett.* **16** 42–44

[62] Haus H A, Fujimoto J G and Ippen E P 1992 Analytic theory of additive pulse and Kerr lens mode-locking *IEEE J. Quantum Electron.* **28** 2086–2096

[63] Blow K J, Doran N J and Nayar B K 1989 Experimental demonstration of optical soliton switching an all-fiber nonlinear Sagnac interferometer *Opt. Lett.* **14** 754–756

[64] Ikeda K 1979 Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system *Opt. Commun.* **30** 257–261

[65] Steinmeyer G, Jaspert D and Mitschke F 1994 Observation of a period-doubling sequence in a nonlinear-optical fiber ring cavity near zero dispersion *Opt. Commun.* **104** 379–384

[66] Haus H A and Islam M N 1985 IEEE *J. Quantum Electron.* **21** 1172–1188

[67] Mollenauer L F and Stolen R H 1984 The soliton laser *Opt. Lett.* **9** 13–15

[68] Haus H A, Fujimoto J G and Ippen E P 1991 Structures for additive pulse mode-locking *J. Opt. Soc. Am.* B **8** 2068–2076

[69] Ippen E P, Haus H A and Liu L Y 1989 Additive pulse mode-locking *J. Opt. Soc. Am.* B **6** 1736–1745

[70] Nakamura S, Ueno Y and Tajima K 1998 Ultrafast (200-fs switching, 1.5-Tb/s demultiplexing) and high-repetition (10 GHz) operations of a polarization-discriminating symmetric Mach–Zehnder all-optical switch *IEEE Photon. Technol. Lett.* **10** 1575–1577

[71] Hirano A, Kobayashi H, Tsuda H, Takahashi R, Asobe M, Sato K and Hagimoto K 1998 10 Gbit/s RZ all-optical discrimination using refined saturable absorber optical gate *Electron. Lett.* **34** 198–199

[72] Whitaker N A, Gabriel M C, Avramopoulos H and Huang A 1991 All-optical, all-fiber circulating shift register with an inverter *Opt. Lett.* **16** 1999–2001

[73] Schiek R 1994 All-optical switching in the directional coupler caused ny nonlinear refraction due to cascaded 2nd-order nonlinearity *Opt. Quantum Electron.* **26** 415–431

[74] Naoum R and Salah-Belkhodja F 1997 Opto-optical switch in nonlinear integrated optics *Pure Appl. Opt.* **6** L33–L36

[75] Ramamurthy P 2002 Ultrafast all-optical switch using LT-GaAs based on DFB *Proc. SPIE* **4643** 266–273

[76] Moores J D, Hall K L, Lepage S M, Rauschenbach K A, Wong W S, Haus H A and Ippen E P 1995 20-GHz optical storage loop laser using amplitude-modulation, filtering, and artificial fast saturable absorption *IEEE Photon. Technol. Lett.* **7** 1096–1098

[77] Moores J D, Wong W S and Hall K L 1995 50-Gbit/s optical pulse storage ring using novel rational-harmonic modulation *Opt. Lett.* **20** 2547–2549

B1.1
Visible light-emitting diodes

Klaus Streubel

B1.1.1 Introduction

Light-emitting semiconductor diodes (LEDs) are light sources that were developed in the last few decades. For most of this time, they have been used as small indicator lights in a wide range of consumer applications. Some 10 years ago, two new material systems, AlGaInP and InGaN, entered the LED arena and gave birth to a new generation of light-emitting diodes: the high-brightness LEDs. This was an important breakthrough for the entire LED business, which enhanced the prospects of LED use in a much wider range of applications. Now, with InGaN covering the emission range from blue to green and AlGaInP from yellow to red, the entire visible spectrum has become accessible to LED light (figure B1.1.1). Furthermore, the continuously improving material quality, together with better chip and package designs, have led to much enhanced performances in terms of efficiency and total output power. Today, at certain wavelengths, LEDs achieve more than 50% energy efficiency in the laboratory and ones with more than 20% efficiency are commercially available. The internal conversion of electrical power into light is sometimes close to 100% and the only task left is to extract as much light as possible out of the semiconductor material without it being lost internally. It can be projected that the good performance at certain colours will eventually be extended to the entire spectrum and the efficiency of commercial LEDs is expected to exceed 50%. Another attractive feature of LEDs is their very long lifetime, of at least some ten thousand hours or several years of continuous operation. Finally, the availability of highly efficient LEDs covering the range from violet to red has now allowed the generation of white light and enabled the LED to enter the wide field of illumination and lighting.

This chapter will give a brief overview of the field of visible light-emitting diodes. The history of the development of visible diodes will be summarized, starting with the early GaAsP and GaP-based devices and continuing up to the development of high-brightness AlGaInP and InGaN LEDs. Next, some basic aspects of the physics of LEDs will be described, with emphasis on their optical and electrical properties. The major semiconductor material systems used for visible LEDs and their fundamental properties are then introduced. Technologies for enhanced light extraction, such as wafer bonding on transparent substrates (TSs), wafer-soldering for substrate-less devices or resonant-cavity and photonic-bandgap designs are discussed next, followed by a description of the most common ways of generating white LED light. Finally, the standard packages for LEDs and some of the most important applications of high-brightness LEDs will be presented. The appendix A includes an introduction into the standard Commission Internationale de L'Eclairage (CIE) colour system, a description of the human eye sensitivity curves and an overview of the most important radiometric and photometric units that are frequently used to describe the properties of LEDs.

At this point, it might be helpful to define some conventions for the following chapter. The most important parameters to characterize an LED are the efficiency and the output power. If not specified

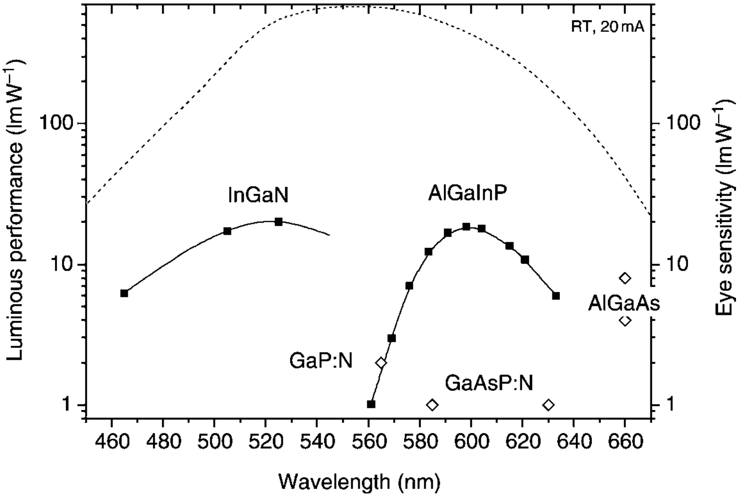


Figure B1.1.1. Luminous performance of commercial high-brightness InGaN and AlGaInP LEDs, compared with a few conventional GaP and GaAsP-based devices. The dotted line shows the human eye sensitivity.

otherwise, efficiency measures the conversion of electrical to optical power and is usually expressed in per cent. For many applications, however, it is more relevant to measure the power of light as perceived by the human eye. In this case, it is convenient to use photometric units such as lumen and candela. The lumen is a measure of the total light power, weighed according to the human eye sensitivity. Efficiency is then expressed in lumens per (electrical)watt. Candelas are a measure of brightness emitted into a certain direction per unit of solid angle and is expressed in lumens per steradian. In the following, the performance of LEDs will be expressed in lumens and lumens per watt (lm W^{-1}) with a few exceptions where mW and per cents are used. Another parameter is the colour of the LED emission. If nothing else is specified, the emission wavelength normally means the peak wavelength of the optical spectrum, which has to be distinguished from the dominant apparent wavelength when the human eye perception is taken into account (see appendix A).

B1.1.1.1 Historical review

In the very beginning of the last century in 1907, H.J. Round discovered a 'curious phenomenon' when he applied a large voltage, of more than 100 V, to a SiC crystal: electroluminescence—the conversion from electrical current to light [1]. Many years later, in the early sixties, electroluminescence was studied extensively on III–V semiconductor alloys such as GaAs or GaAsP. In 1962, the first visible red emitting LEDs were made of GaAsP with a P-fraction of 40%. The devices were fabricated by vapour phase epitaxy on GaAs-substrates, with a pn-junction formed by selective Zn-diffusion into a thick n-type layer. For low P-fractions, GaAsP has a direct energy gap (see section B1.1.3.1) and is almost lattice matched to GaAs. As the P-content is increased, in order to achieve shorter wavelengths, the lattice mismatch and thus the defect density in the material increases appreciably, leading to rather low efficiencies. In 1968, the first commercial red GaAsP LEDs, with luminous efficiencies around 0.2 lm W^{-1}, were introduced to the marketplace by Hewlett Packard. Shortly thereafter, more efficient devices based on GaP doped with Zn and O, were developed using liquid phase epitaxy (LPE) on GaP wafers. GaP:Zn–O LEDs emit at 700 nm, where the human eye sensitivity is down to 0.4% of its maximum value, resulting in luminous efficiencies around $0.4\,\mathrm{lm\,W^{-1}}$. A breakthrough for the early LEDs was the discovery that isoelectronic impurities, such as nitrogen, can act as efficient radiative recombination centres in GaP and GaAsP. GaAsP:N and GaP:N based LEDs already showed efficiencies around $1\,\mathrm{lm\,W^{-1}}$, about one order of magnitude more than the first GaAsP diodes.

In the early 1970s, AlGaAs/GaAs single heterojunction diodes were developed, also using LPE as the method of crystal growth. Starting with emission in the infrared spectral range, the emission wavelength was continuously reduced by increasing the Al-fraction, until red 660 nm emitting diodes were achieved in 1980. The next advance was the introduction of double heterostructures (DHs) for more efficient confinement of the electrical carriers in the active region of the device. The first commercial AlGaAs DH LED with an efficiency around $4\,\mathrm{lm\,W^{-1}}$ was introduced in 1985.

Although more efficient as a result of a better carrier confinement and direct band transitions, AlGaAs/GaAs LEDs had one obvious disadvantage compared to the early GaP-based diodes: unlike GaP, GaAs is not transparent to visible light. The introduction of LPE grown, transparent AlGaAs substrates solved this problem and enhanced the efficiency by almost a factor of two. The first TS LEDs were introduced in 1987 as ‘super-high-brightness devices’ with efficiencies around $8\,\mathrm{lm\,W^{-1}}$.

Because the AlGaAs bandstructure becomes indirect at high Al-fractions, it was not possible to fabricate AlGaAs LEDs emitting at wavelengths shorter than 660 nm. Therefore, a new material system, AlGaInP lattice matched to GaAs, was investigated intensively. Epitaxial growth was only possible using the new method of metal-organic vapour phase epitaxy (MOVPE). The quaternary composition of AlGaInP offered an additional advantage: the bandgap could be tuned from yellow-green to red while maintaining the same lattice constant thus allowing lattice matching on GaAs. With the new material, efficiencies in the order of $10\,\mathrm{lm\,W^{-1}}$ were achieved. As previously with the AlGaAs/GaAs LEDs, the efficiency of AlGaInP/GaAs devices was limited by absorption in the GaAs substrate. A TS AlGaInP LED was therefore developed by Hewlett Packard, using the novel technology of direct wafer bonding. After the growth of the AlGaInP layers on GaAs, the original GaAs substrate was removed and the thin AlGaInP film was transferred (‘wafer-bonded’) to a transparent GaP wafer. As before in the case of TS-AlGaAs LEDs, the efficiency of TS-AlGaInP LEDs was doubled compared to absorbing substrate (AS) LEDs and efficiencies of more than $20\,\mathrm{lm\,W^{-1}}$ were demonstrated on red devices. Another successful method for high efficiency developed at that time was the introduction of thick, transparent window layers giving better current spreading and improved light extraction. In commercial AlGaInP-LEDs, both AlGaAs and GaP windows were employed.

The performance of TS-AlGaInP LEDs was further improved by optimizing the size and optical properties of the chips, yielding efficiencies of more than $70\,\mathrm{lm\,W^{-1}}$ for red emission. Recently, this already impressive performance was further enhanced by shaping the TS-die into truncated inverted pyramids. Record high efficiencies of $>100\,\mathrm{lm\,W^{-1}}$ (610 nm) or >50% (630 nm) were achieved with this technology. Alternatively, Osram-OS started to solder the AlGaInP epitaxial structure onto a new carrier and to remove the original GaAs substrate. The advantage of the wafer-soldering process is its suitability for mass production and its potential to achieve high yields on large diameter wafers. The best ‘substrate-less’ AlGaInP LEDs achieve $>60\,\mathrm{lm\,W^{-1}}$ (615 nm) or up to 40% efficiency (630 nm) with the potential for further improvements.

The early blue emitting LEDs were made of SiC with efficiencies of only $0.1\text{–}0.2\,\mathrm{lm\,W^{-1}}$. In parallel, researchers started to investigate the electroluminescence of GaN-based devices in 1970 [2]. Green and ‘violet’ emitting devices were fabricated but the devices were very inefficient, mainly due to the almost unsolvable problem of p-doping in GaN. The major breakthrough came in 1989, when a Japanese group achieved real p-conductivity by activating Mg dopants with low-energy electron-beam irradiation. Nichia Chemicals finally developed a new technique for efficient Mg-activation by high-temperature annealing and brought the first commercial blue, GaN-based LEDs to the market in 1994.

The GaN was deposited by MOVPE on sapphire substrates. The bandgap of nitride-based LEDs can be tailored over a wide range, from 362 nm (3.4 eV) to 615 nm (2 eV), by adding additional In to the InGaN alloy. This allowed the fabrication of green LEDs and the first blue-green (500 nm) and green (520 nm) devices were also produced by Nichia, for application in traffic lights. As an alternative to the growth on sapphire, which is electrically insulating, InGaN can also be grown epitaxially on SiC wafers. This technology was developed by Cree Research and resulted in the 'superbright' GaN/SiC LED chip introduced in 1995.

With efficient blue and green LEDs at hand to complement the high-brightness red emitters, it became possible to fabricate white LEDs. Several companies developed white LEDs, using a combination of blue, green and red emitting dies in a single package. An alternative approach to generate white light was developed independently in Japan and Germany: the conversion of blue LED light into white using a phosphor wavelength converter. Osram-OS (formerly part of Siemens) developed a phosphor based on $Y_3Al_5O_{12}$ garnet, doped with cerium ions, that absorbed part of the blue emission of a GaN-LED and produced yellow luminescence. The combination of blue and yellow emission produces white light. The phosphor is either suspended in the epoxy resin used for encapsulation or is directly coated on the chip surface. Commercial white LEDs were introduced to the marketplace in 1998 by Osram-OS and Nichia.

B1.1.2 Physics of LEDs

The basic function of an LED is to generate light following the injection of an electrical current into the semiconductor material. The conversion of electrical carriers (electrons) into light (photons) is called electroluminescence and involves two important processes: the excitation of electrons into higher energy states and the relaxation of excited electrons back to empty lower states. If the relaxing electrons release most or all of the energy difference in form of electromagnetic radiation, the process is called a radiative transition. In the case of non-radiative transitions, the energy difference in such a relaxation process is released in form of heat (phonons). The band structure of the semiconductor material plays an important role for the transition processes of excited electrons. In semiconductors with a direct band structure (figure B1.1.2), the excited electrons in the conduction band can relax into states in the valence band under momentum conservation. In materials with an indirect band structure, the transition process requires the assistance of a phonon in order to conserve momentum. Indirect band transitions are therefore much less probable.

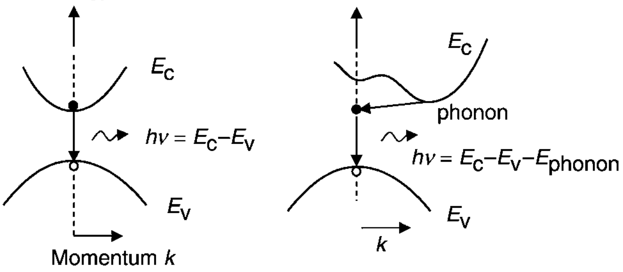


Figure B1.1.2. Left: radiative band–band transitions in a direct bandgap material; right: radiative transition in an indirect bandgap material.

B1.1.2.1 Optical properties of LEDs

Radiative recombination

Figure B1.1.3 shows a number of possible radiative and non-radiative transitions in a simplified band diagram. The direct band–band transition is shown in figure B1.1.3*a*. Because two particles, electrons n and holes p, are involved, this process is called bimolecular recombination. The recombination rate, R (recombinations, per unit volume, per second), is proportional to the density of carriers in the upper state (electrons) and the density of empty lower states (holes) and is given by:

$$R = -\frac{\mathrm{d}n}{\mathrm{d}t} = -\frac{\mathrm{d}p}{\mathrm{d}t} = Bnp \tag{B1.1.1}$$

where n and p are the electron and hole concentrations. The proportionality factor B is a measure of the probability of radiative recombination and is called the bimolecular recombination coefficient. It can be calculated from a few basic material parameters such as the bandgap energy, the absorption coefficient and the refractive index [3] which are all experimentally accessible. For most III–V semiconductors, the bimolecular recombination coefficient is in the order of $1\text{–}10 \times 10^{-10}\,\mathrm{cm^3\,s^{-1}}$. At very high carrier concentrations, as for instance in laser diodes, B starts to decrease and equation (B1.1.1) has to be modified [4].

If the carrier concentrations deviate from the equilibrium, e.g. in the case of electrical or optical excitation, n and p can be written as $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$, and the recombination rate can be split into two parts, an equilibrium recombination rate R_0 and an excess recombination rate ΔR:

$$R_0 + \Delta R = B(n_0 + \Delta n)(p_0 + \Delta p). \tag{B1.1.2}$$

The small term $\Delta n \Delta p$ in equation (B1.1.2) can be neglected. With $R_0 = Bn_0p_0$, the expression can be written as:

$$\frac{\Delta R}{R_0} = \frac{\Delta n}{n_0} + \frac{\Delta p}{p_0} \tag{B1.1.3}$$

with the excess recombination rate given by:

$$\Delta R = B(n_0 + p_0)\Delta n. \tag{B1.1.4}$$

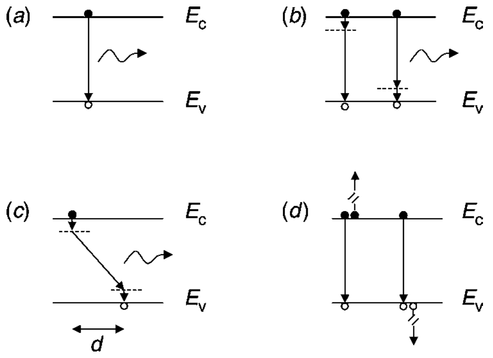


Figure B1.1.3. Recombination mechanisms in a simple band diagram: (*a*) radiative band–band transition; (*b*) band-impurity level transitions; (*c*) donor–acceptor transition; (*d*) Auger recombination process.

In the case of low excitation, with an excess carrier concentration much smaller than $(n_0 + p_0)$, the excess carriers will decay exponentially with time

$$\Delta n(t) = \Delta n_0 \exp\left(-\frac{t}{\tau}\right) \tag{B1.1.5}$$

which gives the time-dependent recombination rate as:

$$\Delta R(t) = -\frac{\mathrm{d}\Delta n(t)}{\mathrm{d}t} = \frac{\Delta n_0}{\tau}\exp\left(-\frac{t}{\tau}\right). \tag{B1.1.6}$$

Comparing equations (B1.1.4) and (B1.1.6), the decay lifetime is obtained as:

$$\tau = \frac{\Delta n}{\Delta R} = \frac{1}{B(n_0 + p_0)} = \frac{1}{R_0}\frac{n_0 p_0}{n_0 + p_0}. \tag{B1.1.7}$$

In the intrinsic case of an undoped semiconductor, with $n_0 = p_0 = n_\mathrm{i}$, equation (B1.1.7) reduces to

$$\tau = \frac{1}{2Bn_\mathrm{i}} = \frac{n_\mathrm{i}}{2R_0}. \tag{B1.1.8}$$

By inserting equation (B1.1.8) in equation (B1.1.1), the radiative recombination rate can be expressed as:

$$R = Bnp = \frac{np}{2n_\mathrm{i}\tau}. \tag{B1.1.9}$$

For n-type materials, the electron is referred to as the majority carrier and the hole as the minority carrier, since $n_0 \gg p_0$. For p-type semiconductors, the roles are reversed. For an n-type semiconductor, we have $n_0 \gg \Delta n \gg p_0$, meaning that the excess carrier concentration is much smaller than the majority carrier concentration, yet much larger than the minority carrier concentration. Because both types of carriers are required for the recombination process, the carrier lifetime that determines the radiative decay of the excess carriers, is the minority carrier lifetime. A very large fraction of majority carriers cannot find a minority carrier for recombination and thus the average majority carrier lifetime is much longer. In the case of doping, the carrier lifetimes become $\tau_\mathrm{n} = 1/(Bp_0)$ and $\tau_\mathrm{p} = 1/(Bn_0)$, for p- and n-doped semiconductors, respectively.

Intermediate states in the semiconductor bandgap may also contribute to the transition from the upper conduction band to the lower valence band, as shown in figure B1.1.3*b*. Examples are transitions from the conduction band to acceptor states or transitions from donor states to the valence band. If the impurity level is deep inside the bandgap, such transitions can be highly probable. Because the impurity has a well-defined position in the crystal lattice, its wave function extends widely in the momentum space. This permits momentum-conserving transitions to occur with reasonable probability even in indirect bandgap materials such as GaP.

Apart from band-impurity level transitions, transitions between two impurity levels are also possible (figure B1.1.3*c*). These donor–acceptor transitions can be rather effective and are very useful for light-emitting diodes. Photons generated in this process have an energy, $h\nu$, given by:

$$h\nu = E_\mathrm{c} - E_\mathrm{v} - (E_\mathrm{D} + E_\mathrm{A}) \tag{B1.1.10}$$

where E_D and E_A are the donor and acceptor binding energies. Coulomb interaction between the donor and acceptor atoms increases the energy of the excited states by an increment of $q^2/\varepsilon d$, where ε is the dielectric constant of the semiconductor host crystal and d the distance between donor and acceptor atoms. This additional energy increases inversely with the separation, d, between donors and acceptors.

Because the spatial distance between donors and acceptors in the host crystal can vary largely, the Coulomb energy increment varies accordingly, causing a wide range of possible photon energies and thus a broad emission spectrum.

Non-radiative recombination

Radiative transitions are electron–hole recombinations in which the excess energy is used to generate a photon. Non-radiative transitions are all other transitions having one feature in common, namely that no photon is generated in the transition process. In fact, in many semiconductors non-radiative carrier recombination is the dominant process. In semiconductors with indirect band structure, such as Si or Ge, the measured radiative lifetime is three orders of magnitude smaller than the calculated value, due to the much larger probability of non-radiative transitions. Experimentally, non-radiative transitions are much more difficult to characterize than radiative transitions, because no characteristic photons are generated. They can be measured only indirectly by analysing the radiative efficiency or the dynamics of the radiation process after external excitation.

The most relevant non-radiative process for optoelectronic devices is recombination via states related to crystal defects. Examples of such defects are dislocations, pores, grain boundaries, vacancies, inclusions or precipitates. Carriers within a diffusion length of the defect will usually be trapped by the defect states and recombine there. The recombination rate scales linearly with the carrier concentration:

$$R_{\mathrm{nr}} = A^{*} n \tag{B1.1.11}$$

where the proportionality factor A^{*} increases with the density of non-radiative centres.

Usually a defect deforms the bandstructure, either by trapping charges or by deforming the lattice, thereby inducing local strain. In either case, the deformation produces a potential barrier of the height, E_{act}, which has to be overcome by the carriers. Therefore, the process recombination has a thermal dependence of the form:

$$R_{\mathrm{nr}}(T) = R^{*} \exp\left(-\frac{E_{\mathrm{act}}}{kT}\right) \tag{B1.1.12}$$

where $R_{\mathrm{nr}}(T)$ is the non-radiative recombination rate and R^{*} a temperature independent coefficient. The temperature dependence of the radiative transition can be neglected. Therefore, as the temperature is reduced, the non-radiative recombination rate decreases exponentially and the radiative processes become more dominant. The temperature dependence of the radiative recombinations over the total number of recombinations is then:

$$\eta_{\mathrm{qi}}(T) = \frac{R_{\mathrm{r}}}{R_{\mathrm{r}} + R_{\mathrm{nr}}} = \frac{1}{1 + \mathrm{z}\dfrac{R^{*}}{R_{\mathrm{r}}} \exp\left(-\dfrac{E_{\mathrm{act}}}{kT}\right)} \tag{B1.1.13}$$

where the ratio R^{*}/R_{r} is constant. The quantity $\eta_{\mathrm{qi}}(T)$ is the internal quantum efficiency.

So far, we have been looking at non-radiative processes inside the semiconductor material. Additionally, minority carriers reaching the surface of the LED are lost due to surface recombination [5]. At the semiconductor surface, the periodic arrangement of atoms, which is the basis for the band structure model, no longer exists. Thus, the surface is a perturbation of the crystal lattice with a strong impact on the band diagram. Due to the lack of neighbours, surface atoms have dangling bonds, partly filled electron orbitals, which can be described as deep or shallow energy levels in the bandgap which may act as recombination centres. The recombination of carriers via surface states is dependent on the semiconductor material and is phenomenologically described by a parameter called the surface recombination velocity, S. It is particularly high in GaAs ($S = 10^6\,\mathrm{cm\,s^{-1}}$) compared to InP ($S = 10^3\,\mathrm{cm\,s^{-1}}$) or Si ($S = 10^1\,\mathrm{cm\,s^{-1}}$). In Al-containing alloys such as AlGaInP, the surface recombination scales appreciably with the Al-fraction [6]. S increases from $10^5\,\mathrm{cm\,s^{-1}}$ for $(\mathrm{Al_{0.1}Ga_{0.9}})_{0.5}\mathrm{In_{0.5}P}$ to $10^6\,\mathrm{cm\,s^{-1}}$ for AlInP.

When designing LEDs, it is very important to consider surface recombination. Carrier injection into the active region of the device should take place several diffusion lengths away from any surface. Recombination at the top-surface of the LED is usually prevented by a high-bandgap confinement layer above the active region and a thick current-spreading or window layer. In some LEDs, light extraction shapes are etched deeply into these top layers. Theoretically, very high extraction efficiencies should be achievable if such shapes would penetrate the pn-junction, but then a substantial number of carriers could recombine at the surface states. However, as long as the active region remains planar, surface recombination can be neglected due to the lack of minority carriers above the confinement layers. In LEDs with a planar surface, carrier diffusion to the lateral edges is reduced by placing the top electrode in the centre of the die.

Instead of generating a photon, the energy released during carrier recombination can also be transferred to other carriers (electrons or holes) and then be dissipated as phonons. This non-radiative process is called Auger recombination. Two examples for Auger processes are shown in figure B1.1.3*d*, but many other processes are possible, depending on the nature and occupation of the involved electronic states. Since two carriers of one type and one carrier of the opposite type are required for the Auger process, the recombination rate is proportional to either np^2 or pn^2. The proportionality factor C is called the Auger coefficient and has typical values for III–V materials of 10^{-28}–$10^{-29}\,\mathrm{cm^6\,s^{-1}}$. It is described as

$$R_{\mathrm{Auger}} = C_{\mathrm{p}}np^2 + C_{\mathrm{n}}pn^2 \tag{B1.1.14}$$

or, in the intrinsic case:

$$R_{\mathrm{Auger}} = Cn_{\mathrm{i}}^3. \tag{B1.1.15}$$

The Auger recombination thereby scales as the cube of the carrier concentration. For LEDs where typical carrier densities are low compared to lasers, the Auger recombination plays a minor role in the device efficiency. However, the Auger effect becomes more prominent in materials with small bandgap energies such as infrared emitting devices.

Emission wavelength

The centre emission wavelength of an LED is mainly given by the bandgap energy $\lambda = hc/E_{\mathrm{g}}$. The full width at half maximum (FWHM) of the emission is approximately $\Delta E \approx 2kT$ [7]. Thus, at a given temperature, $\Delta\lambda/\lambda$ scales with $2kT/E_{\mathrm{g}}$ or $\Delta\lambda$ increases with λ^2. The linewidth can be increased by using high doping levels or having graded compositions in the active region of the LED. The emission linewidth is an important parameter for near-infrared LEDs for optical fibre transmission systems due to the dispersive nature of the fibre.

Light that is extracted through the top surface has to penetrate through only transparent material before reaching the semiconductor–air or –epoxy interface. A significant fraction of light is guided along the active layers and extracted at the die sidewalls. Along this path, the emission wavelength is changed by absorption, which affects mostly the short-wavelength part of the spectrum. This effect is shown in figure B1.1.4 for an AlGaInP LED. Here, the emission spectrum measured at the top-side (0°) differs substantially from the side emission (90°). While the effect is typical for all LEDs that employ band–band transitions, it is not present in LEDs which are based on impurity level transitions, such as

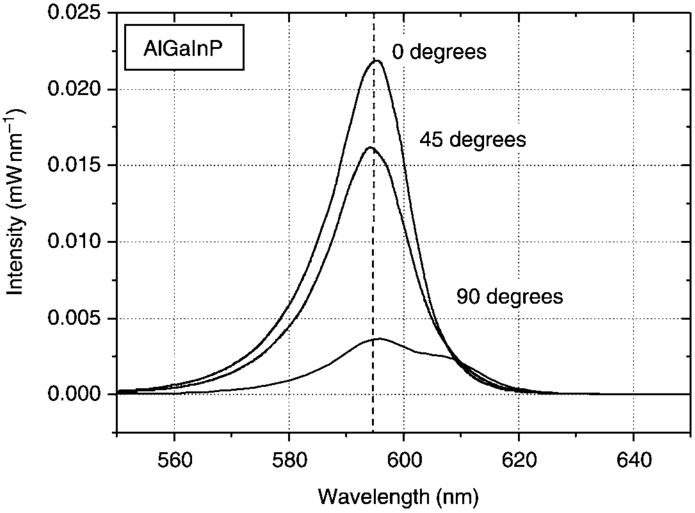


Figure B1.1.4. Spectrum of an AlGaInP LED for different directions of emission. 0° = normal surface emission, 90° = side emission.

used in GaP-based devices. Figure B1.1.14 shows the emission spectra of a GaP:Zn,O LED which is essentially independent of the emission direction.

Light extraction

Although light can be efficiently generated inside an LED ($\eta_{qi} > 90\%$) only a few per cent of the generated photons can actually escape from the semiconductor material. Due to the large difference in the refractive index between the semiconductor ($n = 3-3.5$) and the surrounding medium ($n = 1-1.5$), the interface acts as a perfect mirror for most of the generated photons and prevents them from leaving the diode. The major contribution to the high reflectivity is total internal reflection which occurs at all angles larger than the critical angle ϑ_c. According to Snell's law, this angle is

$$\vartheta_c = \arcsin\left(\frac{n_m}{n_s}\right) \tag{B1.1.16}$$

where n_s and n_m are the refractive indices of the semiconductor and the surrounding medium. All angles are measured to the direction normal to the interface. Only rays of light with $\vartheta < \vartheta_c$ can escape from the diode. In 3D-space this range of angles forms a cone, which is frequently called the 'escape cone'. If the semiconductor chip (e.g. GaAs, $n = 3.5$) is surrounded by air ($n = 1$), the critical angle ϑ_c is 17°, in the case of epoxy encapsulation with a refractive index of $n = 1.5$, ϑ_c becomes 26°. The fraction of escaping photons to the total number of generated photons is then given by

$$\eta_{ext} = \frac{2\pi(1 - \cos\vartheta_c)}{4\pi} = \frac{1}{2}(1 - \cos\vartheta_c) \tag{B1.1.17}$$

which is the ratio of the surface subtended by the spherical escape cone divided by the surface of the full sphere. Even for light that is incident perfectly normal to the chip surface some of the photons are reflected by Fresnel reflection:

$$R_{\mathrm{Fresnel}} = \frac{(n_{\mathrm{m}} - n_{\mathrm{s}})^2}{(n_{\mathrm{m}} + n_{\mathrm{s}})^2}. \tag{B1.1.18}$$

This relation describes the reflection at normal incidence. The correct angular dependence is described by the so-called Fresnel formulae which were first derived by Fresnel in 1823 [8]. For reasons of simplicity we approximate the Fresnel reflection within the light extraction cone by equation (B1.1.18). Then the total extraction efficiency is

$$\eta_{\mathrm{extr}} = \frac{1}{2}(1 - \cos\vartheta_{\mathrm{c}})\left(1 - \frac{(n_{\mathrm{m}} - n_{\mathrm{s}})^2}{(n_{\mathrm{m}} + n_{\mathrm{s}})^2}\right). \tag{B1.1.19}$$

For a GaAs-based LED the fraction of light that escapes the material is as low as ≈ 2% or ≈ 4% for air or epoxy as surrounding medium, respectively. Internal reflection is therefore a major obstacle for the fabrication of high efficiency AlGaInP/GaAs or AlGaAs/GaAs-based LEDs.

The typical destiny of a totally reflected photon is to be reflected a couple of times before finally being absorbed either in the active region or someplace else in the LED structure. Even without the presence of absorption, the angles of reflection in a typically cube-shaped LED die will continue to reproduce themselves, without ever falling inside an escape cone (figure B1.1.5). A straightforward remedy for LEDs on TSs is to give the entire die a geometrical shape which is more suitable for light extraction. The ideal solution would be a spherical or hemispherical chip with small active area in the

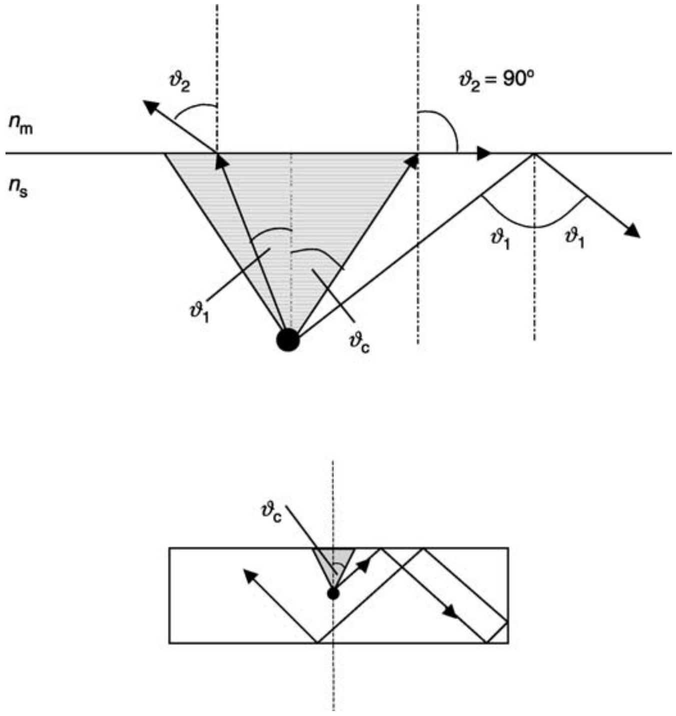
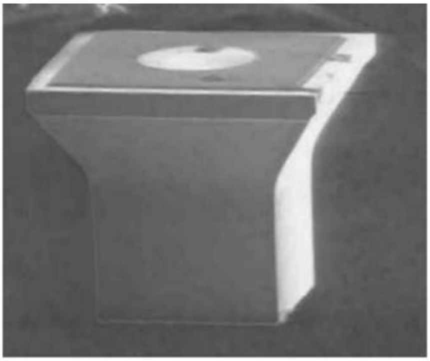


Figure B1.1.5. Top: Total internal reflection and 'escape cone' at the semiconductor–air interface. Bottom: Reproduction of reflection angles in a cubically shaped LED die.

centre. Other geometries are LEDs with hemispherically shaped top side, inverted cones, truncated spheres or truncated cones. Already, in the early sixties, die shaping for high efficiency LEDs was extensively investigated [9, 10]. At that time, the technology of more or less individually shaped dies turned out to be not practical, mostly because of the associated high costs. However, more than 30 years later, high brightness LEDs with geometrically shaped dies have now entered the market: the truncated inverted pyramid AlGaInP LED (see section B1.1.4) and the InGaN/SiC 'Aton' LED, offered by Lumileds and Osram-OS, respectively. A photograph of the Aton-chip is shown in figure B1.1.6. Light is generated in the GaN-based structure on the SiC-substrate. Because the refractive index of SiC is larger than the index of GaN, light rays which enter the SiC material can only cover an angular range of 0–68°, according to Snell's law. Since most of this light would be totally reflected at the SiC–air interface in a cube-shaped die, the side facets are inclined by 30°. This increases the angular range of extracted light rays appreciably and results in an almost doubled brightness compared to the standard rectangular die [11].

From a fabrication point of view, planar LED structures with the shape of a rectangular parallelepiped are favourable. One way to increase the efficiency of planar LEDs is to increase the thickness of the transparent layers above the active layer, until the escape cones parallel to the active layer are fully utilized. This is schematically shown in figure B1.1.7. If the window layer thickness *d* is



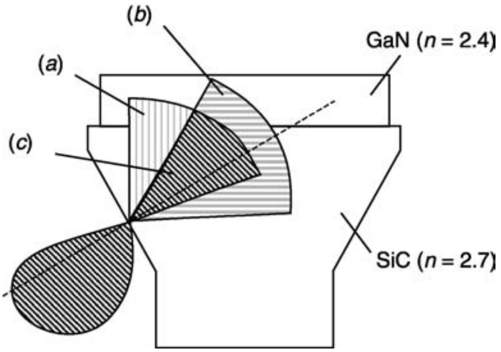


Figure B1.1.6. Above: InGaN 'Aton' chip with inclined facets. Below: Principle of light extraction in the 'Aton'-die: (*a*) range of angles coupled from GaN into SiC, (*b*) range of angles that can transmit the SiC–air interface, (*c*) fraction of light in SiC that can be extracted.

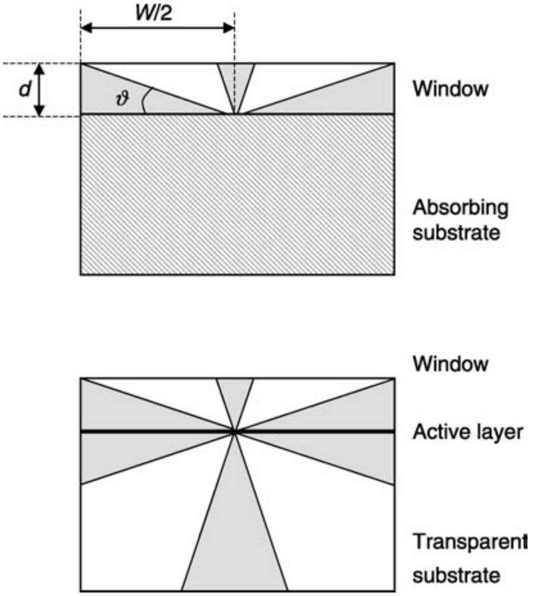


Figure B1.1.7. Escape cones in a cubic LED die. Top: LED with AS and thick window layer. Bottom: TS LED.

big enough, light generated in the centre and emitted into the side escape cone can escape if:

$$d_{\text{window}} = \frac{w}{2} \tan \vartheta_{\text{c}} \tag{B1.1.20}$$

where w is the lateral size of the die. For an encapsulated GaP LED with $n_{\text{s}} = 3.4$ and $n_{\text{m}} = 1.5$, the critical angle ϑ_{c} is 26°. If the chip width is $w = 300$ μm, the optimum window layer would be $d_{\text{window}} = 74$ μm. If the substrate is transparent, it is theoretically possible to extract the escape cones from all six sides of the cube, which would give 24% extraction efficiency for an encapsulated LED. In case of an AS, the bottom cone and the bottom half of the four side cones are lost, leaving an extraction efficiency of 12%, if the window layer complies with equation (B1.1.20). This is of course only a very crude model to estimate the effect of transparent windows and substrates because important effects like multiple pass reflection, absorption and re-emission are neglected. Real LEDs with TS and thick window layer can achieve up to 32% external efficiency [12] exceeding even the theoretical value of 24% using the simplified model.

Bragg mirror—About half of the emission of an LED is directed towards the substrate. If the substrate is not transparent, this light is basically lost by absorption resulting in a substantial reduction in extraction efficiency. This loss can be partly reduced by the introduction of a Bragg mirror between the active layer and the AS. A Bragg mirror consists of a stack of layer pairs with different refractive indices, where each layer has the thickness of a quarter optical wavelength. In such a stratified layer structure, each interface provides a certain reflectivity, depending on the contrast in refractive index. The Fresnel reflectivity at each interface is given by:

$$R_{\text{Fresnel}} = \frac{(n_{\text{low}} - n_{\text{high}})^2}{(n_{\text{low}} + n_{\text{high}})^2}. \tag{B1.1.21}$$

Here, n_{low} and n_{high} are the refractive indices of the layer with higher and lower index, respectively. Because all layers have the thickness of a quarter wavelength, all reflections in that direction add in phase so that the total reflection for perpendicular incidence can be very high. If the number of layer pairs is m, the mirror reflectivity is:

$$R_{\mathrm{DBR}} = \left[\frac{1 - \left(\frac{n_{\mathrm{low}}}{n_{\mathrm{high}}} \right)^{2m}}{1 + \left(\frac{n_{\mathrm{low}}}{n_{\mathrm{high}}} \right)^{2m}} \right]^2 . \qquad \text{(B1.1.22)}$$

This shows that a high mirror reflectivity can be achieved even for low index contrast materials, if the number of layer-pairs m is high enough. However, besides the practical drawback of growing large number of layer pairs, a low index contrast also results in a very narrow reflection bandwidth both in terms of wavelength and angular range. Therefore, Bragg mirrors are only of practical use if the material system offers a ratio $n_{\mathrm{high}}/n_{\mathrm{low}}$ sufficiently larger than 1.

In an LED, the Bragg mirror should be tuned to a maximum power reflectivity integrated over a range of angles within the escape cone. Note, that this situation is typically not achieved with a Bragg mirror optimized for perpendicular incidence. By shifting the centre reflectivity of a DBR towards longer wavelength, the reflectivity for off-axis light rays is enhanced and, since the solid angle with a given angular increment in these directions is much larger, a significantly higher integral reflectivity is obtained. If we take a 630 nm AlGaInP-LED, with a 20 period $Al_{0.5}Ga_{0.5}As/AlAs$ DBR, the integral reflectivity for a centre wavelength at 630 nm is in the order of 50%, whereas it is close to 75% if the mirror is tuned to 670 nm.

Surface structuring—Another approach to enhance the extraction efficiency of LEDs on ASs is to texture the surface [13]. Figure B1.1.8 shows a schematic drawing of a commercial surface-structured high-brightness AlGaInP-LED. The light extraction structure is etched into the window layers above the pn-junction. The basic idea is to inject the electrical current only along the edges of the die so that most of the light is generated close to the edge. The surface structure is then optimized for the extraction of light generated below the electrical contacts. An additional benefit of this approach is that the light is generated far away from the bond pad, which would otherwise shadow a substantial fraction of the emission. Other approaches for high extraction efficiency are discussed in section B1.1.4.

Radiation patterns

If external factors such as reflection, absorption or re-emission are excluded, the spontaneously generated photons in an electronic transition are emitted isotropically into all directions. The radiant

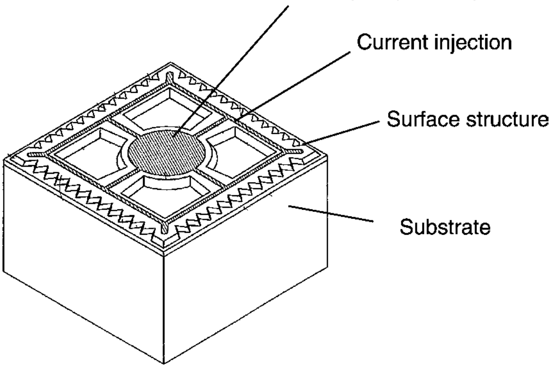


Figure B1.1.8. Sketch of a surface structured LED, with an etched light-extraction structure along the contacts at the chip edges.

intensity I_e of an LED is the divergence of optical power at a defined angular direction. In the case of ideal isotropic emission, the radiant intensity I_e would have the same value of $I_e = P_{out}/\Omega = P/4\pi$ in all directions. The solid angle Ω of a full sphere has the value $\Omega = 4\pi$. In the non-isotropic case, I_e is defined as the integral of $dP_{out}/d\Omega$ over all angles. Ideal isotropic radiation can only be achieved with point light sources and without the presence of internal reflection. In a real semiconductor diode, however, light is generated in an active region with the lateral extension of the die and is extracted through the semiconductor–air interface. In this case, the radiant intensity can be described by a cosine dependence:

$$I_e(\vartheta) \approx I_0 \cos(\vartheta). \tag{B1.1.23}$$

This is the Lambertian cosine law of radiation which applies to most planar LEDs. LEDs with TSs such as GaP- and AlGaAs/AlGaAs-LEDs or wafer-bonded AlGaInP/GaP devices, have enhanced side emission. Some novel types of LEDs show totally different emission patterns. Examples are LEDs with shaped dies such as the truncated inverted pyramid LED [14] or the 'Aton LED' [15]. Other examples are resonant-cavity LEDs, where the radiation pattern is given by the built-in optical resonator [16] (figure B1.1.9). Besides the properties of the LED die, the radiation characteristics of encapsulated LEDs is strongly influenced by the used package.

B1.1.2.2 Electrical properties

Electroluminescence

In a semiconductor diode, an electrical current composed of a flow of electrons and holes, is injected into the active region. A forward bias across the pn-junction reduces the potential barriers for electrons and enables them to spill over into the p-region (figure B1.1.10). Similarly, the forward bias reduces the potential barrier which blocked the flow of holes into the n-region. The overlapping concentration of electrons and holes allows spontaneous recombination.

The most simple diode structure is a pn-homojunction LED, where the p- and n-doped regions of a semiconductor material form a pn-junction. Examples of homojunction LEDs are blue GaN-LEDs where the pn-junction is created during the eptiaxial growth of n- and p-type GaN layers. Better carrier confinement is achieved, if the pn-junction is formed by two semiconductors having different bandgaps. Single heterojunction (SH) LEDs typically involve an n-type semiconductor with a narrow bandgap and a p-type semiconductor with a wider bandgap. SH active regions were first employed for LEDs in the AlGaAs system. Even better injection efficiency of electrons and holes and improved carrier confinement can be achieved if two heterojunctions are used to form the active region. Double heterostructure active regions of LEDs use a narrow bandgap semiconductor, sandwiched between two wide bandgap materials (figure B1.1.11) [17, 18]. In semiconductor lasers, the narrow bandgap layer has a thickness of a few ten nanometres, whereas in LEDs it can be up to some micrometres thick. High-brightness AlGaInP- and InGaN-LEDs employ double heterostructure active regions.

If A_{pn} and d_{pn} are the area and thickness of the active region, the generation density of carriers is

$$G = \frac{I_a/q}{A_{pn} d_{pn}} = \frac{J_a}{q d_{pn}} \tag{B1.1.24}$$

with $J_a = I_a/A_{pn}$ as the current density. If the generation rate of carriers equals the recombination rate in the active layer of the LED, the system has reached an equilibrium:

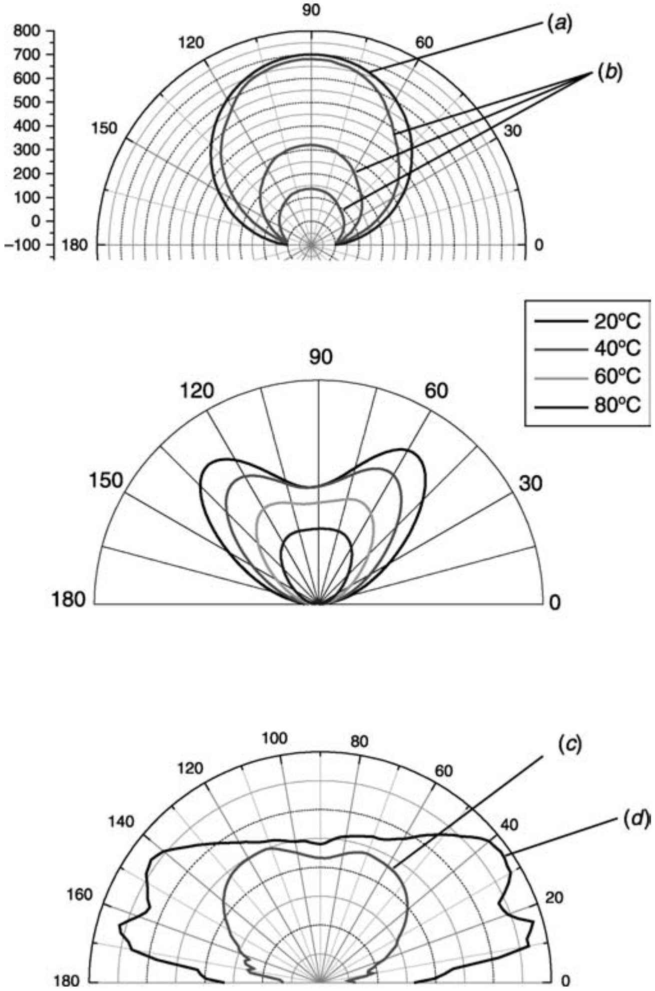


Figure B1.1.9. Emission pattern of different LED dies: top: ideal Lambertian radiator (*a*), AlGaInP LED at different current (*b*), middle: 650 nm resonant cavity LED, bottom: conventional SiC/InGaN LED (*c*) and InGaN 'Aton' die (*d*).

$$G = \frac{J_\mathrm{a}}{q d_\mathrm{pn}} = An + Bn^2 + Cn^3. \tag{B1.1.25}$$

The right-hand side of equation B1.1.25 is the total recombination rate for a given density of injected electrons n. The coefficient A in the linear term accounts for surface recombination, as well as recombination via defect levels. For typical, undoped III–V materials, A is in the order of $10^8\,\mathrm{s}^{-1}$. The quadratic term includes the bimolecular recombination coefficient B and accounts for radiative transitions. The cubic term is the Auger recombination.

According to equation (B1.1.25), the radiative recombination rate increases with the square of the carrier density, whereas the non-radiative recombination is basically a linear function of the carrier density (neglecting Auger). As the carrier concentration in the active region increases, the radiative

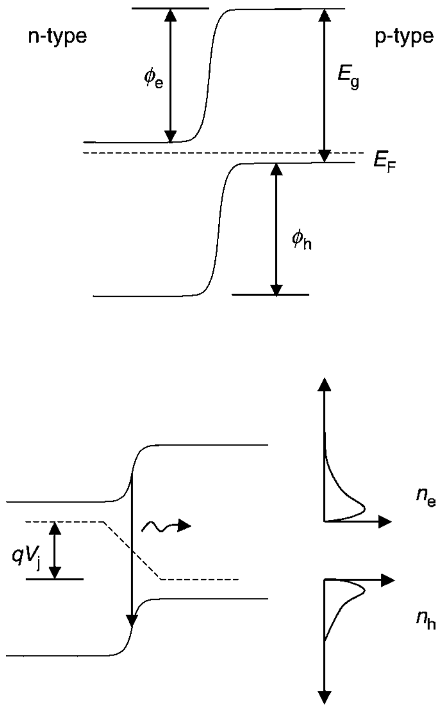


Figure B1.1.10. Band structure at the pn-junction of a simple homostructure LED. Above: with no bias and below: under forward bias. Lower right: electron and hole distributions.

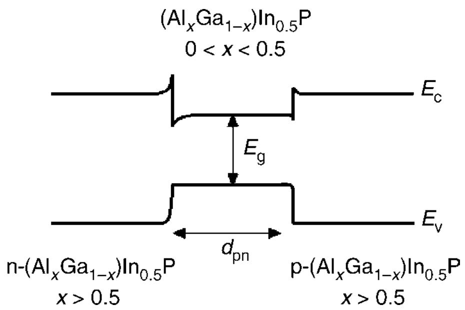

$(Al_xGa_{1-x})In_{0.5}P$
$0 < x < 0.5$
E_c
E_g
E_v
d_{pn}
$n\text{-}(Al_xGa_{1-x})In_{0.5}P$
$x > 0.5$
$p\text{-}(Al_xGa_{1-x})In_{0.5}P$
$x > 0.5$

Figure B1.1.11. Schematic diagram of a double heterostructure LED active layer in the AlGaInP-system.

process starts to dominate over the non-radiative process and the internal efficiency of the device will increase. High carrier density in the active layer is usually achieved with a double heterostructure, where the active layer is sandwiched between large-bandgap barrier layers. The number of carriers that spill over the barriers is exponentially dependent on the barrier height, as well as on the temperature. Thus, to minimize this carrier leakage, the heterobarriers should be as high as possible, at least a few kT. While some alloy systems, such as AlGaAs/GaAs or InGaN/GaN, offer the possibility to realize high potential barriers, some other systems, like AlGaInP, lack large direct bandgap compositions.

The typical ternary and quaternary III–V alloys used for LEDs consist of solid solutions of binaries with direct (GaAs, InP) and indirect (GaP, AlP) bandgaps. As the fraction of indirect materials is increased, the bandstructure can change from direct to indirect. The closer the composition of an active region is to the so-called 'cross-over' point, where the indirect conduction band minimum (X) becomes lower than the direct minimum (Γ), the more carriers are leaking into the X-minimum. The Γ–X transfer of carriers is a non-radiative loss mechanism which substantially limits the internal efficiency of yellow and green emitting devices (see section B1.1.3).

Current–voltage characteristics

The current–voltage characteristic of pn-junctions is described by the Shockley equation [19, 20], which is given by:

$$I(V) = I_\mathrm{s}\left(\exp\frac{qV}{kT} - 1\right) \tag{B1.1.26}$$

where I_s is the current obtained with a reverse bias (V negative). I_s is called the saturation current and is controlled by the number of minority carriers which diffuse to the pn-junction. In this (ideal) case, the current density, I_s/A, can be calculated according to:

$$\frac{I_\mathrm{s}}{A} = \frac{qD_\mathrm{p}p_\mathrm{n0}}{(D_\mathrm{p}\tau_\mathrm{p})^{1/2}} + \frac{qD_\mathrm{n}n_\mathrm{p0}}{(D_\mathrm{p}\tau_\mathrm{n})^{1/2}} \tag{B1.1.27}$$

where A is the cross-sectional area of the diode junction, D_n and D_p are the electron and hole diffusion constants, n_p0 and p_n0 are the equilibrium electron and hole concentration on the p- and n-side, respectively. Under forward bias for $V > 3kT/q$ the current rises exponentially with V.

To describe real diodes, with a series resistance R_s, and in the presence of recombination centres in the active material, the Shockley equation has to be modified:

$$I(V) = I_\mathrm{s}\left(\exp\frac{qV - qR_\mathrm{s}I}{\alpha kT} - 1\right). \tag{B1.1.28}$$

The factor α is the so-called ideality factor. In the case of an ideal diode, α becomes 1, and, in the case of recombination only via defect states in the depletion region, α becomes 2. Since in real diodes both processes take place, the ideality factor α will take a value between 1 and 2. The series resistance R_s is typically of the order of a few ohms and can be neglected at small currents <1 mA. However, at higher currents R_s has an important influence on the $I(V)$ characteristics and on the energy efficiency.

A typical current–voltage characteristics of an AlGaInP-LED at room temperature is shown in figure B1.1.12. A linear current scale is used for the upper half of figure B1.1.12, whereas the lower part uses a logarithmic scale. At currents in the microampere region and below, non-radiative transitions dominate the recombination mechanisms and the ideality factor α is 2. For currents between approximately one microampere and one milliampere, more carriers are injected into the active region and the non-radiative transitions become saturated with carriers. At those currents, the desired radiative recombination can compete successfully with the transitions via non-radiative centres and the ideality factor α reduces to 1.3–1.5. At currents above some milliamperes, the series resistance of the LED eventually determines the slope of the current–voltage characteristics.

B1.1.2.3 Efficiencies

In an ideal LED, every injected electron generates one photon with an energy of $h\nu = E_\mathrm{g}$. In this case, the number of injected carriers I/q equals the number of generated photons and the efficiency of this process is 1. However, in reality, not all injected electrons reach the active region, not all electrons in the active

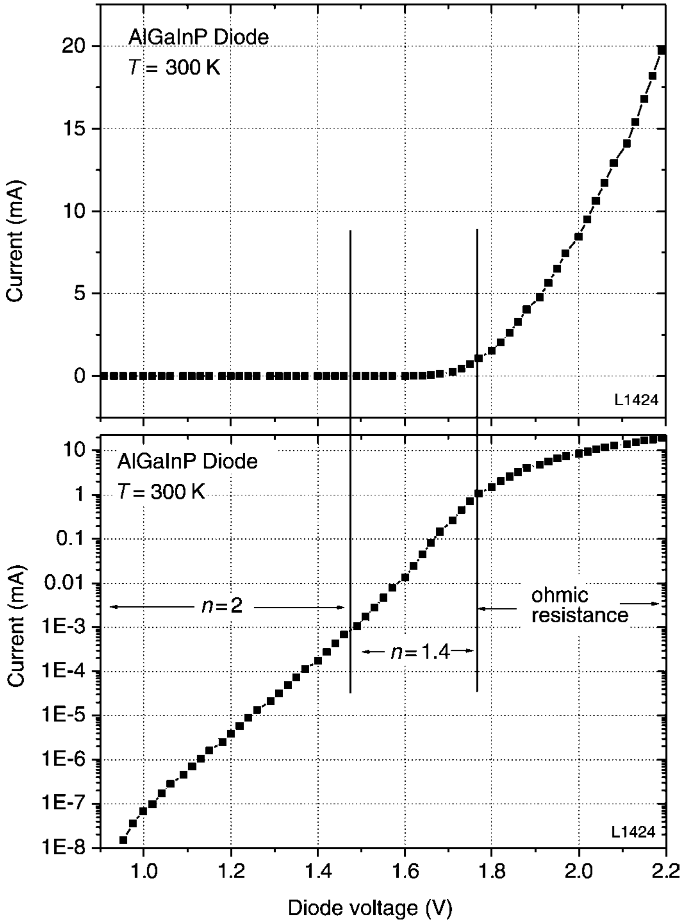


Figure B1.1.12. Current–voltage characteristics of an AlGaInP diode at room temperature on a linear (top curve) and logarithmic (bottom curve) current scale.

region can generate photons and not all photons can escape from the semiconductor. Therefore, all involved processes like carrier injection, carrier recombination or photon extraction are more or less efficient. Their efficiencies can be expressed either as quantum efficiency in terms of the number of photons and electrons involved or as power efficiency in terms of the optical or electrical power of all carriers involved.

In real LEDs, some of the electrical current is lost, e.g. due to conductive channels along the surfaces of the die or at crystal defects in the vicinity of the pn-junction. The fraction of current which is actually reaching the active region is:

$$I_a = \eta_{inj} I_0. \tag{B1.1.29}$$

The ratio $\eta_{inj} = I_a/I_0$ is called the injection efficiency.

At the pn-junction, the carriers recombine, either via radiative or non-radiative transitions. The internal quantum efficiency η_{qi} is defined as the number of generated photons per unit time over the number of electrons injected per unit time. In some cases, it is more convenient to express η_{qi} in terms of radiative and non-radiative recombination rates (equation (B1.1.13)) or the corresponding carrier lifetimes

$$\eta_{\mathrm{qi}} = \frac{R_{\mathrm{r}}}{R_{\mathrm{r}} + R_{\mathrm{nr}}} = \frac{\tau_{\mathrm{nr}}}{\tau_{\mathrm{r}} + \tau_{\mathrm{nr}}}. \tag{B1.1.30}$$

Only photons that leave the semiconductor chip are useful. Since only a small portion of photons generated in the die can be extracted, we have to define an extraction efficiency $\eta_{\mathrm{extraction}}$ as the number of extracted photons over the total number of generated photons.

The external quantum efficiency of an LED is the product of the internal quantum efficiency and the extraction efficiency:

$$\eta_{\mathrm{ext}} = \eta_{\mathrm{qit}}\eta_{\mathrm{extraction}}. \tag{B1.1.31}$$

This is the ratio of the number of photons emitted from the LED per unit time to the number of injected electrons into the LED per unit time. The number of emitted photons from the LED is obtained by dividing the measured optical energy by the photon energy $h\nu$. The number of electrons per unit time is the injected electrical current divided by the elementary electron charge q. Thus, the external quantum efficiency, η_{ext}, can be calculated according to

$$\eta_{\mathrm{ext}} = \frac{\left(\dfrac{P_{\mathrm{out}}}{h\nu}\right)}{\left(\dfrac{I_0}{q}\right)} = \frac{P_{\mathrm{out}}q}{h\nu I_0}. \tag{B1.1.32}$$

In an ideal diode, the forward voltage, V_{f}, of the device equals the bandgap energy divided by the elementary electron charge E_{g}/q. In real LEDs, however, the device structure contains additional series resistances and potential barriers which will increase the forward voltage. Therefore, we can define an external power efficiency or ‘wall-plug efficiency’ as

$$\eta_{\mathrm{wp}} = \frac{P_{\mathrm{out}}}{I_0 V_{\mathrm{f}}}. \tag{B1.1.33}$$

The ‘wall-plug’ efficiency is the most important parameter for LED applications. It can be expressed in terms of several efficiencies defined above:

$$\eta_{\mathrm{wp}} = \eta_{\mathrm{inj}}\eta_{\mathrm{qi}}\eta_{\mathrm{extraction}}\frac{h\nu}{qV_{\mathrm{f}}}. \tag{B1.1.34}$$

The different definitions are summarized in table B1.1.1.

B1.1.3 Material systems for visible LEDs

B1.1.3.1 GaP and GaAsP

The evolution of ternary III–V alloys started with the development of epitaxial growth methods like LPE or VPE. Binary alloys like GaAs or GaP were first available as bulk crystals and could be used as substrate materials. If P is added to an epitaxial layer of GaAs, the emission is shifted from the infrared to the visible spectrum (figure B1.1.13) and the lattice constant becomes larger than that of GaAs. For low P-fractions up to about 0.45 in $\mathrm{GaAs}_{1-x}\mathrm{P}_x$, the bandstructure remains direct and direct red emission is achieved [21]. However, the more P is added to the alloy, the larger the lattice mismatch to

Table B1.1.1. Summary of different definitions of internal and external efficiencies. The last expression for the extraction efficiency is valid for LEDs with a planar structure.

	Quantum efficiency	Power efficiency
Internal efficiency	$\eta_{\mathrm{qi}} = \dfrac{\#\,\text{of generated photons}}{\#\,\text{of injected electrons}}$	$\eta_{\mathrm{pi}} = \dfrac{\text{power of one photon}}{\text{power of one electron}}$
	$\eta_{\mathrm{qi}} = \dfrac{R_{\mathrm{r}}}{R_{\mathrm{r}} + R_{\mathrm{nr}}} = \dfrac{\tau_{\mathrm{nr}}}{\tau_{\mathrm{r}} + \tau_{\mathrm{nr}}}$	$\eta_{\mathrm{pi}} = \dfrac{h\nu}{qV}$
External efficiency	$\eta_{\mathrm{ext}} = \dfrac{\#\,\text{of extracted photons}}{\#\,\text{of injected electrons}}$	$\eta_{\mathrm{wp}} = \dfrac{\text{extracted optical power}}{\text{injected electrical power}}$
	$\eta_{\mathrm{ext}} = \dfrac{P_{\mathrm{out}} q}{h\nu I_0}$	$\eta_{\mathrm{wp}} = \dfrac{P_{\mathrm{out}}}{I_0 V_{\mathrm{f}}} = \eta_{\mathrm{inj}} \eta_{\mathrm{qi}} \eta_{\mathrm{extraction}} \dfrac{h\nu}{qV_{\mathrm{f}}}$
Injection efficiency	$\eta_{\mathrm{inj}} = \dfrac{\#\,\text{of carriers at pn junction}}{\#\,\text{of injected carriers}} = \dfrac{I_{\mathrm{a}}}{I_0}$	
Extraction efficiency	$\eta_{\mathrm{extraction}} = \dfrac{\#\,\text{of extracted photons}}{\#\,\text{of generated photons}}$	
	$\eta_{\mathrm{extraction}} = \frac{1}{2}(1 - \cos\vartheta_{\mathrm{c}})\left(1 - \dfrac{(n_{\mathrm{m}} - n_{\mathrm{s}})^2}{(n_{\mathrm{m}} + n_{\mathrm{s}})^2}\right)$	

I_0, total injected current; I_{a}, current flow into the pn-junction; R_{rad}, radiative recombination rate; R_{nr}, non-radiative recombination rate; τ_{r}, radiative carrier lifetime; τ_{nr}, non-radiative carrier lifetime; P_{out}, optical power emitted from the LED; V, applied voltage; ϑ_{c}, angle of total reflection; n_{s}, refractive index of the semiconductor material; n_{m}, refractive index of the surrounding medium; #, number.

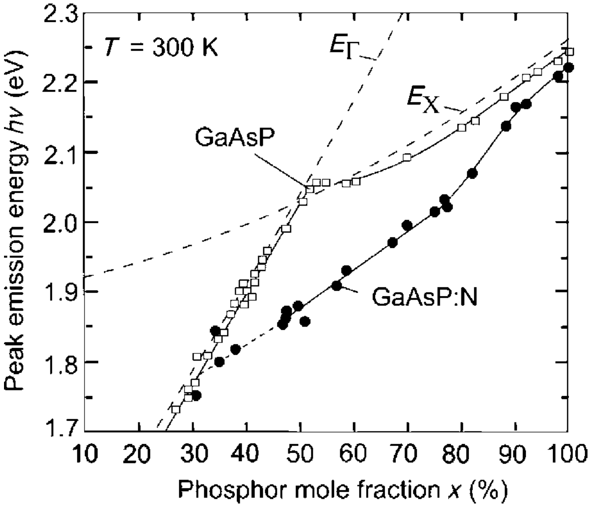


Figure B1.1.13. Emission energy versus phosphorous mole fraction of GaAsP and GaAsP:N LEDs. The solid lines are obtained from fits to experimental data at $5\,\mathrm{A\,cm^{-2}}$. The dashed lines are calculated data for the direct and indirect bandgap [21].

GaAs becomes and the more misfit dislocations are generated. The external efficiency of a GaAsP LED decreases by about a factor of 10 as the P-fraction is increased to the direct–indirect crossover point at $x = 0.45$. Fortunately, the human eye response increases in this wavelength range by more than a factor of 50 so that the product of LED efficiency and eye sensitivity reaches a maximum at $x = 0.4$ ($\approx$ 650 nm, red). The fabrication process of these LEDs is rather simple. It includes a selective Zn-diffusion step into an n-type layer to form the pn-junction and subsequent contact formation.

The emission properties are significantly changed if the material is doped with an isoelectronic impurity such as nitrogen. As a group V-element, nitrogen does not produce additional carriers as standard dopants do, but introduces a carrier trapping level approximately 100 meV below the conduction band. Trapped electrons become localized in space which causes a large spread in the allowed momentum of the electron. Matching the momentum of holes in the valence band enhances dramatically the probability for radiative recombination and thus the generation of light. The problem of lattice-mismatch-induced dislocations is, of course, still present and limits the efficiency for higher P-fractions in GaAsP [22]. On the other hand, isoelectronically doped LEDs do not suffer from re-absorption to the same extent as intrinsic devices, which is a substantial advantage for higher efficiency. Consequently, the light extracted in different directions from the die does not show the variation in the wavelength spectrum that is observed in high-brightness LEDs (figure B1.1.14).

Doping GaP simultaneously with the two group V-elements N and O also enhances the efficiency of GaP-LEDs. Introduced together in similar quantities, the two dopants tend to form complexes after proper annealing, and these function as isoelectronic traps. The Zn–O complex level lies deeper in the GaP band gap (approximately 0.3 eV below the conduction band) than the N-level and produces red photons at a wavelength of 700 nm. The low sensitivity of the human eye at 700 nm and the rapid saturation of the Zn–O centres with increasing current density limits the luminous efficiency to values below $0.5\,\mathrm{lm\,W^{-1}}$.

Table B1.1.2 summarizes some characteristic features of the most common GaAsP and GaP-based LEDs. One property of transitions via impurity levels is, that the recombination rate saturates as more carriers are injected, which limits the maximum applicable current to the diode. The brightness of

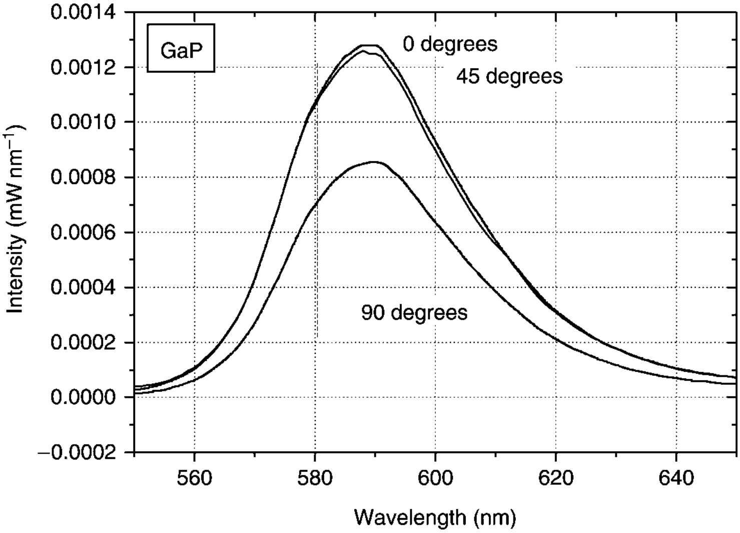


Figure B1.1.14. Emission spectrum from a red GaP:Zn,O LED showing a directionally independent emission spectrum.

Table B1.1.2. Properties of the most common types of GaAsP- and GaP-based LEDs.

	Substrate	Colour	Peak wavelength (nm)	Ext. efficiency (%)	Performance ($\mathrm{lm\,W^{-1}}$)
$GaAs_{0.6}P$	GaAs	Red	650	0.2	0.2
$GaAs_{0.35}P_{0.65}$:N	GaAs	Orange–red	630	0.7	1
$GaAs_{0.14}P_{0.86}$:N	GaAs	Yellow	585	0.2	1
GaP:N	GaP	Green	565	0.4	2.5
GaP:Zn–O	GaP	Red	700	2	0.4

isoelectronically doped LEDs, such as green GaP:N devices, is still sufficient for many applications such as indoor signs, indicators or illumination of small displays. However, although such LEDs are still sold in quantities of billions of devices per year, the performance is not sufficient for high brightness applications.

B1.1.3.2 AlGaAs/GaAs

The first high-brightness LEDs on the market were red $Al_xGa_{1-x}As/GaAs$ devices. The ternary alloy $Al_xGa_{1-x}As$ is a solid solution of AlAs ($E_g = 2.168\,\mathrm{eV}$) and GaAs ($E_g = 1.424\,\mathrm{eV}$). Since AlAs ($a = 5.66\,\mathrm{\AA}$) and GaAs ($a = 5.65\,\mathrm{\AA}$) have basically the same lattice constant, AlGaAs is conveniently lattice matched to a GaAs substrate for the entire composition range. While GaAs is a direct bandgap material, AlAs has an indirect bandstructure. The bandgap of AlGaAs is direct for an Al-fraction up to 45% and indirect for higher values. This limits the efficient emission of visible light to the red spectral range above 620 nm. The composition dependence of the Γ- (direct gap) and the X-minima (indirect gap) of the Brillouin zone can be calculated according to reference [23]:

$$E_{g\Gamma}(x) = 1.424 + 0.247x \ \ \mathrm{eV} \quad 0 \le x \le 0.45 \tag{B1.1.35}$$

$$E_{g\Gamma}(x) = 1.424 + 0.247x + 1.147(x - 0.45)^2 \ \ \mathrm{eV} \quad 0.45 \le x \le 1 \tag{B1.1.36}$$

$$E_{gX}(x) = 1.9 + 0.125x + 0.143x^2 \ \ \mathrm{eV} \quad 0 \le x \le 1.0. \tag{B1.1.37}$$

The energies of the AlGaAs conduction band minima of the Γ-, X- and L-bands, are shown in figure B1.1.15 as a function of the Al-mole fraction. The band discontinuity at an AlGaAs/AlGaAs heterojunction splits in a ratio of roughly 60:40 between the conduction and valence bands [24]. Typical doping elements for p-type AlGaAs are zinc, carbon, beryllium and magnesium, while n-doping can be achieved with silicon, selenium or tellurium. The electrical conductivity of the material decreases rapidly with increasing Al-content.

Today, AlGaAs is a mature material system, readily fabricated using LPE or MOVPE growth [25]. High brightness visible LEDs emitting between 650 and 660 nm are routinely produced and are used worldwide. High performance AlGaAs-LEDs utilize an active layer with a double heterostructure confinement layer. Grown on absorbing GaAs substrates, these devices achieve luminous performances around $3\,\mathrm{lm\,W^{-1}}$. The material system offers the possibility to grow thick (100–150 μm) epi-layers by LPE on GaAs substrates. With these very thick AlGaAs layers, it is possible to remove the original GaAs substrate by, e.g. selective etching, leaving a transparent ‘substrate’ LED structure. Such TS AlGaAs LEDs are roughly twice as bright as devices on GaAs substrates with luminous performances of $8\,\mathrm{lm\,W^{-1}}$. One major drawback of AlGaAs is its tendency to oxidize in the presence of oxygen or

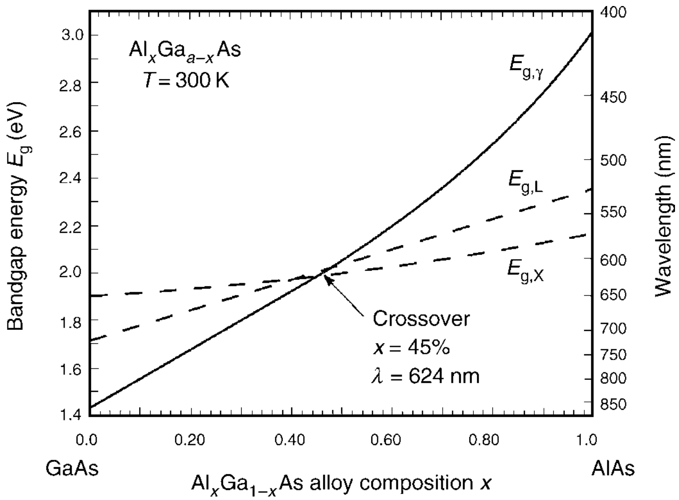


Figure B1.1.15. Composition dependence of the AlGaAs conduction band minima at the Γ-, X- and L-point.

moisture. Because the standard epoxy materials cannot seal the LED die efficiently, the oxidation is also present after polymer encapsulation. The process is accelerated at elevated temperatures. Although visible AlGaAs LEDs are still used in large quantities, it is anticipated that they will be replaced in many outdoor applications by the more robust AlGaInP LEDs.

B1.1.3.3 AlGaInP/GaAs

The alloy system $(Al_xGa_{1-x})_yIn_{1-y}P$ is a direct semiconductor material that covers the range of colours from red (650 nm) to yellow-green (560 nm). For higher brightness applications, it has widely replaced the indirect bandgap emitters GaP:N and GaAsP:N. The quaternary alloy AlGaInP can be regarded as a mixture of three binary materials, AlP, GaP and InP. The first two binaries have nearly the same lattice constant, which is appreciably different from the third involved material InP (figure B1.1.16). In order to obtain lattice matching to GaAs, the In mole fraction is fixed at about 0.48, whereas the Al mole fraction x can be varied widely without affecting the lattice constant (figure B1.1.17).

As usual for zinc blende type lattices, the band structure of the material has a single conduction band, with minima at the Γ-, X- and L-points, and three valence bands with degenerate heavy and light hole bands at the Γ-point. For low Al-content the bandgap is direct, but becomes indirect, with X being the lowest conduction band level for increasing Al-content. The bandgap energies have been determined by various methods [26], yielding slightly different results. A commonly accepted relation for the bandgap variation with composition at room temperature is:

$$E_{g\Gamma}(x) = (1.900 + 0.61x)\ \text{eV} \tag{B1.1.38}$$

$$E_{gX}(x) = (2.204 + 0.085x)\ \text{eV}. \tag{B1.1.39}$$

These values indicate that the Γ–X crossover takes place for $x = 0.58$, corresponding to an energy of 2.25 eV or a wavelength of 550 nm. As the crossover composition is approached from the low-energy side, the radiative efficiency of the material decreases drastically due to transfer of electrons from the Γ- to the X-valley, determining the maximum accessible range of wavelengths for LEDs. The conduction L-valley has not been seen directly in any experiment, but from measurements involving hydrostatic

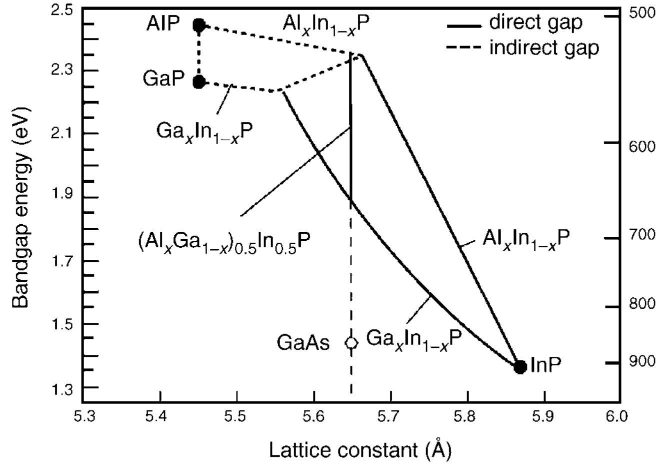


Figure B1.1.16. Bandgap versus lattice constant for the quaternary AlGaInP-system.

pressure [26], it has been estimated to be at least 125 meV above Γ_c. Although some theoretical work indicates a closer proximity to Γ_c [27], there are negligible effects on the performance of luminescent devices.

Various data indicate that the common 60:40 rule for the conduction and valence band offsets is also applicable in AlGaInP material [28]. The X-minimum of the conduction band changes accordingly, leading to a maximum band offset of about 200 meV in the conduction band. This conclusion has been nicely confirmed by transport experiments on n–i–n heterostructures [29].

For the active material, a double heterostructure design or a multiple quantum well (MQW) structure, embedded between larger bandgap layers, is used to optimize the carrier confinement. As the Al-concentration in the active region is increased to achieve shorter wavelength emission, electrons are thermally transferred from the Γ-minimum to the X-minimum, generating an additional non-radiative

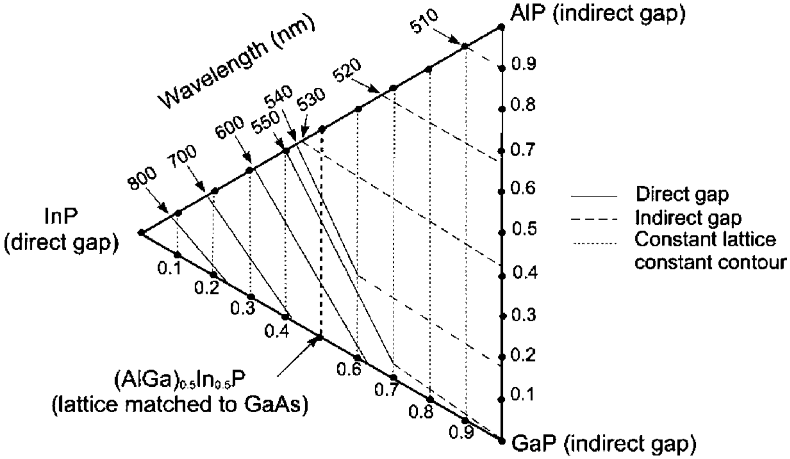

Wavelength (nm)
510
520
530
540
550
600
700
800
AlP (indirect gap)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
GaP (indirect gap)
InP
(direct gap)
0.1
0.2
0.3
0.4
0.6
0.7
0.8
0.9
$(AlGa)_{0.5}In_{0.5}P$
(lattice matched to GaAs)
Direct gap
Indirect gap
Constant lattice constant contour

Figure B1.1.17. Two-dimensional representation of the AlGaInP composition ranges. Solid lines are lines of constant wavelength, dotted lines mark compositions of constant lattice parameter. Some indirect bandgap wavelengths are shown as dashed lines.

recombination channel. Beyond the Γ–X-crossover, the radiative efficiency drops to zero. Closed-form expressions for the non-radiative losses through Γ–X-transfer have been derived under certain approximations [28] but a fit to experimental data requires additional assumptions, specifically a reduction of carrier lifetimes with Al-concentration. While this assumption is reasonable (the residual oxygen incorporation, e.g. increases with the Al content), electron leakage and, to a minor extent, hole leakage across the barriers follow a similar relationship on current and temperature, making it difficult to distinguish between the mechanisms in simplified models.

Simulations of the device efficiency show that, in the red colour range, the internal quantum efficiency is solely determined by non-radiative processes [30]. For emission wavelengths below 600 nm on the other hand, carrier leakage supersedes the non-radiative losses. Carrier leakage can be minimized by using larger active layers. Lower carrier densities in the active layer lead to higher effective barriers. At the same time, however, non-radiative processes become more favourable. Thus, as in many cases, the optimum configuration is determined by the material quality and different results have been obtained by different research groups [12, 13]. Another way of suppressing the carrier leakage is to increase the barrier heights. The barriers are maximized by using doped AlInP for the confinement layers. Even larger barriers can be obtained by thin layers of AlInP, with tensile strain or multiple quantum barrier (MQB) structures [31]. In the latter case, a combination of barrier layer and short-period strained superlattice (SPSS) is used, such that the lowest SPSS miniband accessible for electrons tunneling out of the active layer is above the largest conduction band energy of any of the constitutant materials.

An important issue for the fabrication of optoelectronic devices is the tendency of the material to form ordered phases. Depending on the growth conditions, atomic ordering of In and Al/Ga along the (111) planes can occur on the group-III sublattice [32]. The modified crystal structure leads to a significant reduction of the bandgap energy, which has been measured to be up to 160 meV in partially ordered material [33] and is estimated to be as large as 471 meV for completely ordered material [34]. Moreover, in photoluminescence experiments, it has been found that peaks can be split and broadened [35] and luminescence intensities are decreased [36], possibly due to the formation of piezoelectric fields and non-radiative recombination centres at domain boundaries. Therefore, disordered material is more favourable for most LED devices, in particular when short wavelengths are needed. Generally, the use of misoriented substrates, e.g. (100) tilted towards the $\langle 110\rangle$ or the $\langle 111\rangle_A$ direction, and the proper choice of growth conditions and doping profiles allows the complete suppression of ordering in the material [37]. More details about the AlGaInP alloy system can be found in the literature [28, 30, 38].

B1.1.3.4 InGaN

The group-III nitrides are the only III–V system that allows the generation of UV, blue and green photons. (Al)InGaN alloys can be described as a solution of the three binaries AlN, GaN and InN, with a very restricted solid miscibility due to thermodynamic constraints [40]. All three materials are direct semiconductors with band gaps of 6.2, 3.4 and 1.9 eV, respectively [41]. By adding In to GaN, the emission wavelength is shifted from the UV (365 nm) to the visible spectrum. Theoretically, the (Al)InGaN-system can cover a wide wavelength range up to 630 nm. The practical limit today, however, is somewhere in the green spectral range (≈ 550 nm).

The alloys can crystallize in two different modifications, either in a cubic zinc blende structure or in a hexagonal wurzite structure with the latter being considered as thermodynamically more stable. Since neither GaN nor AlN are available as bulk crystals, the system lacks a suitable semiconductor substrate for epitaxial growth (figure B1.1.18). The most common substrate today is sapphire, with a lattice mismatch of 16% to GaN. Although this large mismatch causes a very high dislocation density between 10^7 and $10^9\,\mathrm{cm}^{-2}$ in the epitaxial films [42], the system is still capable of producing highly efficient LEDs,

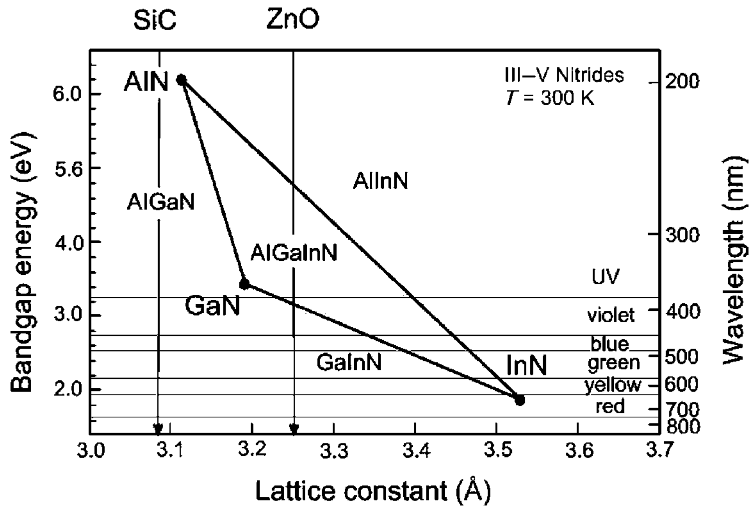


Figure B1.1.18. Bandgap energy versus lattice constant for III–V alloy systems of InGaN. The bowing parameters for the energy bands of AlGaN and InGaN are assumed to be 1.0 according to [39].

for reasons that are not fully understood. Obviously, the dislocations that must occur do not act as efficient recombination centres, which, together with the low diffusion length of holes, do not appreciably affect the radiative recombination efficiency. An alternative substrate is SiC which has a lower lattice mismatch to GaN and offers a number of additional advantages such as good electrical and thermal conductivity or a high electrostatic discharge hardness [43, 44]. Compared to sapphire, SiC is about ten times more expensive and less transparent. Also, its refractive index is less suitable for light extraction.

It has been found that the growth of a GaN, AlN or AlGaN buffer layer is essential for the growth of high-quality layers on both mismatched substrates SiC and sapphire. This buffer is usually grown at low temperatures (500–600°C), before the temperature is raised to 1000–1100°C for the growth of the device structure. On sapphire substrates, the buffer layer is necessary to achieve a smooth surface morphology and good optical and electrical properties despite the large lattice mismatch [45–47]. On SiC-substrate, the buffer layer reduces the voltage drop across the SiC–GaN interface. Also, the buffer thickness directly impacts the quality of the subsequently grown layers. Maximum Hall mobility in n-type GaN is achieved with 20 nm thick GaN-buffer layer thickness [47]. From the different temperature dependence of the Hall mobility with GaN- and AlN-buffer layers, it might be concluded that the GaN-buffer reduces the effect of ionized impurity scattering [48].

Nominally undoped GaN is n-type, with a high residual electron concentration. P-doping is very difficult to achieve in III-N-alloys. The lack of p-type material, and hence of pn-junctions, has been a blocking point for the fabrication of light-emitting devices for a long time. A lot of the early work was done with Zn as acceptor [49, 50], but today the preferred material for p-doping is magnesium [51]. After the epitaxial growth, the acceptors are passivated by the formation of some sort of Mg–H complexes [52]. P-conductivity can be obtained after a post-growth treatment by thermal annealing or using low-energy electron irradiation (LEEBI), but also ‘in situ’ by a suitable growth and cooling-down procedure. The activation energy of Mg acceptors in GaN, deduced from temperature-dependent Hall measurements, is assumed to be around 160 meV. In InGaN and AlGaN-alloys, the activation of Mg

Table B1.1.3. Material properties of binary III-nitrides and their potential substrates. GaAs and Si are listed as reference. Note that the thermal conductivity of SiC is anisotropic. The value in brackets gives the value in a different crystallographic direction.

Material	Lattice constant (Å)	Thermal expansion coefficient ($10^{-6}\,\mathrm{K}^{-1}$)	Thermal conductivity ($\mathrm{W\,cm^{-1}\,K^{-1}}$)	Refractive index
GaN	$a = 3.160 \cdots 3.190$[a]	5.59	1.3	2.29 @ 1000 nm[a]
	$c = 5.125 \cdots 5.190$	7.75		
InN	$a = 3.5446$[a]		–	2.56 @ 500 nm[a]
	$c = 5.7034$			
AlN	$a = 3.11$[a]	5.27[a]	2.0	
	$c = 4.98$	4.15		
6H-SiC	$a = 3.0806$[a]	4.9	4.89 (0.878)	$2.55378 + 3.417 \times 10^4\ \lambda^{2}$[a]
	$c = 15.1173$			
Sapphire	$a = 4.758$[b]	7.70 (⊥ c)	0.439	ordinary ray: 1.78 @ 589 nm[c]
				(extraordinary ray: −0.008)
		8.33 (‖c)		
GaAs	5.65325	5.8	0.46	3.5 @ 600 nm
Si	5.43102	2.29	1.48	3.94 @ 600 nm[d]

[a] Landolt Börnstein data handbook, new series.
[b] S. Nakamura, Semiconductor and Semimetals, vol 48.
[c] E.D. Palik, Handbook of Optical Constants of Solids II.
[d] E.D. Palik, Handbook of Optical Constants of Solids I.

acceptors is dependent on the composition. Silicon or germanium is used for n-type doping. The donor activation energy of Si in GaN is around 33 meV.

As a consequence of the high ionicity of the nitrides, their refractive indices are relatively small. The values for the binaries are listed in table B1.1.3. The index of (Al)InGaN can be changed appreciably by changing the composition and thus the bandgap. The refractive index is an important parameter for light extraction from the diode chip. Another favourable property of the III-nitrides is their high thermal conductivity. Some of the values are listed in table B1.1.3. The common substrates such as sapphire and 6H-SiC have thermal conductivities of 0.439 and 4.9 $\mathrm{W\,cm^{-1}\,K^{-1}}$, respectively.

The first blue emitting LEDs were homotype pn-junction GaN devices. The electroluminescence spectrum consists of a narrow acceptor-band emission in the UV (around 370 nm) and a broad band-acceptor line at 420 nm. Better carrier confinement and thus more efficient recombination are achieved in double heterostructure devices. By adding Al to GaN, the bandgap is increased while the addition of In shifts the bandgap to lower energies. Thus, typical DH-devices involve AlGaN confinement layers around GaN or InGaN active layers. Due to the large difference of lattice constant between AlN and InN, the range of usable composition and layer thickness is limited. Similar to the impurity related transitions in GaP, simultaneous doping of InGaN with Zn and Si can generate relatively efficient donor–acceptor emission in the violet to green spectral range [53, 54]. Using active material of quantum wells with a well thickness below the critical thickness the quality of the LED structure is improved significantly, which enables the use of band–band or near band–edge transitions as major recombination process [55]. As long as mismatch-induced strain and the quantum well thickness do not exceed the critical limits, the emission wavelength can be readily adjusted with the In-mole fraction. Nakamura *et al* reported emission from violet to yellow for In-fractions between 0.15 and 0.7 [56].

B1.1.4 High efficiency LEDs and novel technologies

The traditional design of an LED chip is a semiconductor die, with a planar metal contact at the bottom and a circular or square electrode at the top (figure B1.1.19). Very simple devices use a homostructure pn-junction as active region, the more advanced ones include a single- or double-heterostructure. Several factors limit the external efficiency of such LEDs: total internal reflection, shadowing by the metal electrodes and absorption in the semiconductor material. With the exception of GaP-based LEDs and some AlGaAs-devices on TSs, a substantial amount of light is lost by absorption in the substrate material. Therefore, even with an efficient internal process of current-light conversion, the external efficiency of standard LEDs hardly exceeds a few per cent. This frustrating situation has stimulated a lot of work on ways to increase the efficiency of LEDs such as:

- fabrication of dices that are as transparent as possible, by the use of very thick transparent layers and a TS;
- die-shaping to enhance the extraction efficiency;
- fabrication of substrate-less LEDs, to avoid substrate absorption;
- modification of the spontaneous emission by using optical cavities.

Each approach tackles the problem of light extraction in a different way. The first one applies rigorously the principle of extracting light from as many escape cones as possible. Die shaping is a step towards the ideal solution, that of a point light source in the centre of a spherical semiconductor die. Most of such ideas are almost as old as the LED itself, but practical devices have not been generated until very recently. For obvious reasons, die shaping works best on LEDs with TSs. The (absorbing) substrate can be removed entirely, if the LED layers are transferred to a new carrier such as a semiconductor wafer or other materials such as metal or ceramic. If metal soldering is used to combine the LED structure with the carrier, the metal also acts as mirror. Because the reflected light has to pass through the entire layer structure, the efficiency of LEDs with such an integrated reflector is very sensitive to internal absorption (table B1.1.4).

A completely different idea for high extraction efficiency is to change the fundamental process of spontaneous emission. Placing light generating material in an optical resonator, with spatial dimensions in the order of the emission wavelength, can change the spontaneous generation of light significantly

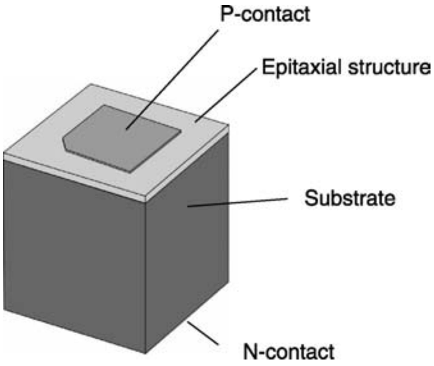


Figure B1.1.19. Conventional LED die on a conductive substrate. Typical dimensions are a chip area of $200 \times 200\ \mu m^2$ and a height of $200\ \mu m$.

Table B1.1.4. Summary of LEDs with the highest reported efficiencies. Listed are the material system, the technology used, the colour range of emission and the status (productive or demonstrator). The different LEDs are described in more detail in the text.

Device	Material system	Technology	Colour range (nm)	Status
Transparent substrate LED	AlGaInP/GaAs	Direct wafer bonding	590–650	Prod.
Truncated inverted pyramid LED	AlGaInP/GaAs	Direct wafer bonding, die shaping	590–650	Prod.
Thin-film LED	AlGaInP/GaAs	Wafer soldering	590–650	Prod.
Surface textured thin-film LED	AlGaAs/GaAs	Epitaxial lift-off, surface roughness	850–980	Demo.
LED with tapered waveguide	AlGaAs/GaAs	Etched lateral taper	850–980	Demo.
	InGaAlP/GaAs		650	Demo.
Resonant Cavity LED	AlGaAs	Standard technology	850–980	Demo.
	AlGaInP/GaAs		650	Prod.

[57, 58]. The ultimate micro-cavity LED would allow only one single optical resonator mode which could be extracted as efficiently as the optical modes in a laser. While such ideal LEDs are still far from any practicably manufacturable device, some aspects of micro-cavity LEDs have already found their way into commercial LEDs, e.g. red emitting (650 nm) RCLEDs.

B1.1.4.1 Transparent substrate AlGaInP LEDs

The extraction efficiency of LEDs with TSs is substantially higher than that of LEDs on ASs. Examples of LEDs on TSs are the early GaP-devices or AlGaAs-LEDs fabricated on thick LPE grown transparent layers. Due to the large lattice mismatch to GaP, high-brightness AlGaInP must be grown on absorbing GaAs-wafers, which strongly reduces the extraction efficiency. In order to overcome this fundamental limitation, Kish *et al* developed a technique to remove the absorbing GaAs substrate and to transfer the AlGaInP-LED layers to a transparent GaP wafer [59]. Direct wafer bonding allows the combination of various semiconductor materials more or less independently of their crystallographic lattice constant. In the case of AlGaInP–GaP, wafers with up to 3 inch diameter are bonded [60], the limit set by the diameter of the available GaP-wafers. The clean surfaces of both wafers are brought into contact under uniaxial pressure and heated up to temperatures of 750°C or more [61]. Under such conditions, and with proper crystallographic alignment, the two materials form a semiconductor heterointerface with covalent bonds between the two materials [60]. The interface is optically transparent and conducts both heat and electrical current. Excellent control over this complex process is required, in order to ensure low-resistance electrical contact over the entire bonded area. The feasibility of the AlGaInP–GaP wafer bonding process for the fabrication of an LED has been demonstrated by Hewlett-Packard and is now routinely used in high-volume production. The new class of AlGaInP LEDs is named TS LEDs to distinguish them from AS LED on GaAs-wafers.

With the very thick GaP current spreading layer on top of the AlGaInP active region and the transparent GaP-wafer below, the TS-LEDs achieve record high levels of light extraction. Very high efficiencies of up to 32% at 630 nm have been reported [12]. In terms of luminous efficiency, a maximum value of 74 lm W^{-1} has been achieved at shorter wavelengths (615 nm). The extraction efficiency of TS-LEDs can be further enhanced, if the die is shaped into cones, pyramids or spheres [62]. The highest reported efficiencies of AlGaInP LEDs are achieved with wafer-bonded TS-material which is cut into dices with the shape of a truncated inverted pyramid [14]. Light is generated at the base of the inverted pyramid and is extracted after a reduced number of reflections and with a low average photon path length within the semiconductor. The truncated inverted pyramid LED achieves a luminous efficiency of 102 and 68 lm W^{-1} at wavelengths of 610 and 598 nm, respectively. A peak external efficiency of 55% was measured at 650 nm.

B1.1.4.2 Wafer-bonded thin-film LED

A viable alternative to direct wafer bonding is to solder wafers with an intermediate metal layer. This is well established in silicon technology and is commonly used for the fabrication of micro mechanical components. In the context of AlGaInP-LEDs, the idea is to transfer the epitaxial structure at wafer level to a new carrier by soldering followed by GaAs removal [63]. This process generates a new wafer, with a metal layer buried between the epitaxial LED layers and the carrier material. Both the high reflectivity of the metal–semiconductor interface and the possibility to form ohmic contacts are favourable for the functionality of the LED.

The process flow is shown schematically in figure B1.1.20. After epitaxial growth, Au and AuSn are deposited on the AlGaInP-LED structure and the carrier wafer, respectively. Carrier and LED-structure are then brought into contact and soldered at 350°C. Then the original GaAs substrate is removed by selective wet chemical etching. Processing of the 'new wafer' is then finished using standard LED processing technology.

Figure B1.1.21 shows the basic principle of a thin-film LED. The metal layer below the active layer serves both as a reflector and as the anode. Light which is not directly extracted is reflected either by total reflection at the semiconductor–air (or epoxy) interface or at the metal mirror. Extraction is significantly enhanced if the reproduction of reflection angles is suppressed by surface roughness. Other critical parameters for high efficiency are a sufficiently thin active layer for reduced self-absorption and a power reflectivity of more than 90% at the metal–semiconductor interface. Although metals like Au, Al or Ag offer a very high reflectivity on AlGaInP or AlGaAs, the reflectivity is greatly reduced if the metals are alloyed to form good electrical contacts. Even on highly doped semiconductors contact alloying is

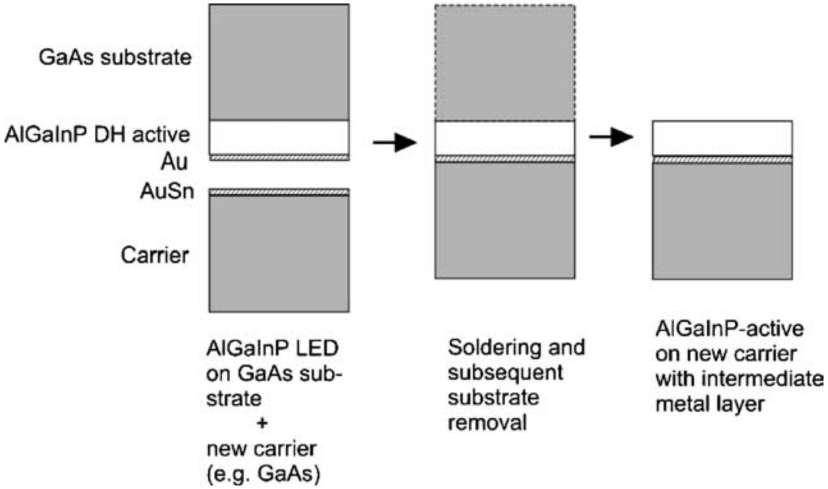

GaAs substrate
AlGaInP DH active
Au
AuSn
Carrier
AlGaInP LED on GaAs sub-strate + new carrier (e.g. GaAs)
Soldering and subsequent substrate removal
AlGaInP-active on new carrier with intermediate metal layer

Figure B1.1.20. Schematic layout of the thin-film process sequence. Left: Metal layers of Au and AuSn are deposited on the epitaxial AlGaInP-structure and the carrier wafer, respectively. After bonding and GaAs removal, the LED structure together with the new carrier forms an artificial wafer (right) which can be processed using standard LED processing.

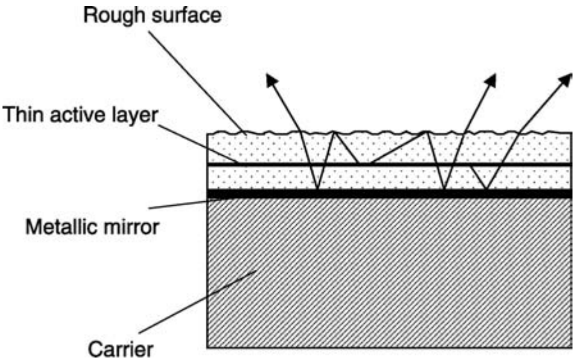


Figure B1.1.21. Principle of operation of an LED with metallic reflector.

necessary to reduce the contact resistance. For the thin-film LED, this problem is solved by locally separating areas with high reflective mirrors from areas with good electrical contacts to the semiconductor. A thin dielectric layer, such as Si_2N_3 or SiO_2, is deposited between the semiconductor and the metal layer(s). Openings in this dielectric film define the electrical contacts, whereas the rest of the area serves as a dielectric/metallic mirror. The layer system can be alloyed in order to lower the contact resistance, without significantly changing the optical properties of the mirror area.

An attractive feature of wafer soldering for LEDs is the possibility to structure one of the surfaces before bonding. Geometrical shapes, such as cones, prisms or spheres, might be transferred into the AlGaInP-structure (or the carrier surface) by etching before they are covered with metal and embedded inside the LED structure. Generally, most of the geometries that have been used to shape dices, in order to enhance the extraction of light [62], can also be buried at the bonded interface of a thin-film LED. An example for an LED with an embedded array of cone-shaped micro reflectors is shown in figure B1.1.22. Truncated cones are etched through the active layer and electrically contacted at openings in the dielectric layer at the soldered interface. The fact that the localized current injection leads to a preferred generation of light in the centre of the etched cones is used to design the actual shape of the microstructure. Light rays that are totally reflected at the metal mirror along the cones are directed upwards towards the LED surface where they are extracted. The path length from the active layer to the top of the LED structure is only a few micrometre and does not include any absorbing material such as the active layer so that the effect of self-absorption inside the device is greatly reduced.

Figure B1.1.23 (left) shows the top view of a 615 nm device with a structured interface. Although the array of micro reflectors is located underneath the planar top-layers, it is clearly visible from the top.

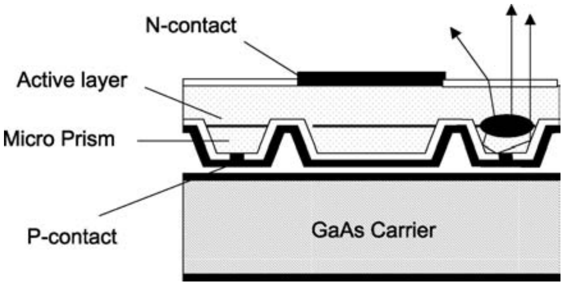


Figure B1.1.22. Schematic cross-section of a thin-film LED with buried micro reflectors.

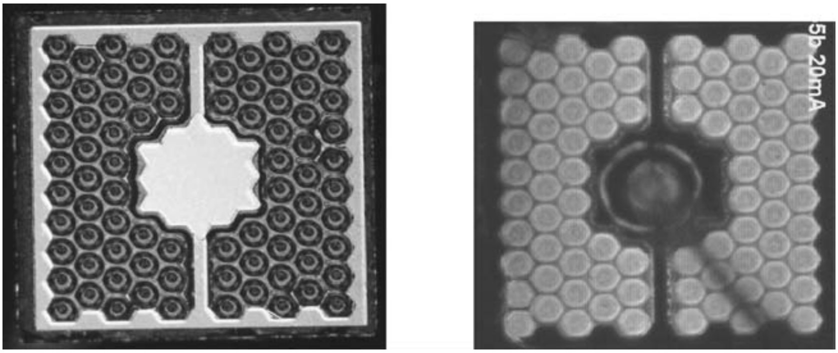


Figure B1.1.23. Top view (left) and illumination pattern (right) of a 615 nm thin-film LED.

Current injection is facilitated from the isolated bond pad and the connected square frame along the edge via the top n-contact layer to the individual p-electrodes in the centre of the individual reflectors. As shown in the illumination pattern (figure B1.1.23, right) light is generated and extracted around the micro reflectors, confirming the principle of operation. Figure B1.1.24 depicts the optical characteristics of the device. Driven at a dc-current of 100 mA, the LED emits up to 9.7 lm, corresponding to an optical power around 32 mW. In the range of 10–20 mA of operation current, the luminous performance is above 50 lm W^{-1}, with a peak value of 53 lm W^{-1} at 10 mA. One of the advantages of the thin-film technology is the low ohmic resistance that can be achieved. Operated at a dc-current of 10 mA, the forward voltage of the 615 nm LEDs is still below 2.0 V.

Thin-film LEDs are also very attractive as large-area chips for high-current application. Contrary to most other high-brightness LEDs, they do not need a thick window layer for light extraction. Because the thickness of these GaP or AlGaAs window layers is directly related to the chip size (equation (B1.1.20)), an increase of chip area usually results in a reduced extraction efficiency. The output characteristics of two large-area thin-film LEDs with 700 and 1000 μm chips are shown in figure B1.1.25. Up to a drive current of several hundred mA, the efficiency of both devices is comparable. Only at very high currents, the output power of the smaller chip begins to roll over due to thermal problems. The 1 mm LED achieves 64 lm at 1 A.

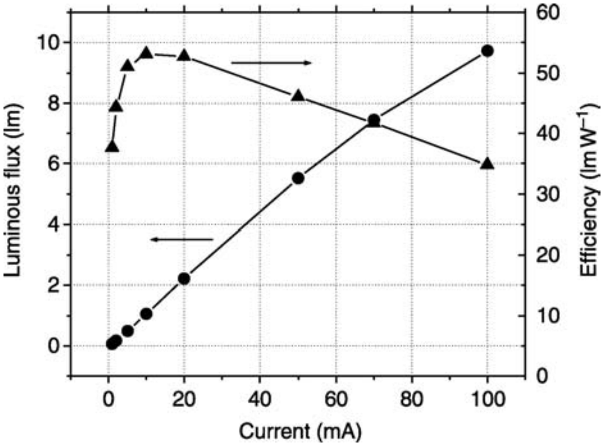


Figure B1.1.24. Optical performance of a 615 nm thin-film LED operated under dc-conditions.

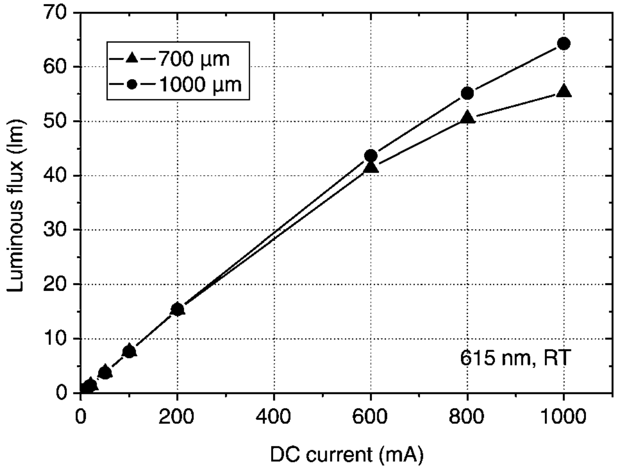


Figure B1.1.25. Light-output characteristics of a 1 mm thin-film chip emitting at 615 nm. The samples are mounted in a package for high-current operation.

The possibility of shaping the LED surface before bonding offers a whole range of new opportunities to optimize or tailor the device performance by combining novel micro-structures for light extraction with new schemes for current injection, heat dissipation or emission profiles. Figure B1.1.26 shows the performance of a 637 nm thin-film LED with a modified reflector structure. At 50 mA, the device emits a power of 34 mW corresponding to a quantum efficiency of 36%. The maximum values of quantum and wall-plug efficiency are 39.5 and 38%, achieved at a drive current of 10 and 2 mA, respectively.

A different approach to fabricate thin-film AlGaInP-LEDs is to bond the epitaxial structure to an isolating wafer. In this case, two top electrodes are required. Horng *et al* fabricated 600–620 nm AlGaInP LEDs on Si-wafers coated with SiO_2 [64, 65] using a metal combination of Au and AuBe for bonding. Despite the intermediate dielectric layer, the LEDs benefited from the good thermal properties

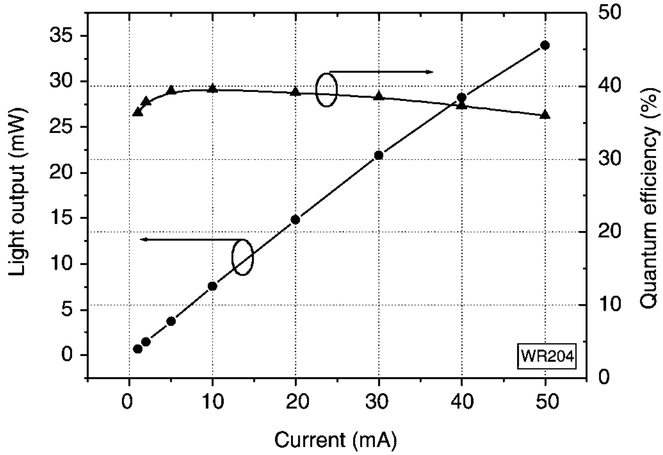


Figure B1.1.26. Characteristics of a 637 nm thin-film LED. The device gives 34 mW of output power at 50 mA dc-current. Maximum values of quantum- and wall-plug efficiency are 40% (10 mA) and 37.7% (2 mA), respectively.

of Si. A luminous intensity of 90 and 205 mcd (620 nm) was demonstrated at 20 and 50 mA of operation current, respectively.

Obviously, the thin-film approach is not restricted to a specific material system. Recently, a 460 nm InGaN thin-film LED was reported achieving an external quantum efficiency of 25% [66]. The device structure was grown on a sapphire and subsequently bonded to a GaAs wafer. A laser lift-off process was used to remove the original sapphire substrate.

B1.1.4.3 Surface-textured thin-film LED

The idea of surface-textured thin-film LEDs dates back to the early 1990s, when Schnitzer *et al* demonstrated an external efficiency as high as 72% with an optically pumped thin-film LED [67] and, shortly later, an electrically driven diode with an external efficiency of approximately 30% [68]. The devices are based on two new techniques for LEDs: 'epitaxial lift-off' (ELO) [69] and surface roughening by 'natural lithography' [70]. The idea of the ELO process is to insert a sacrificial layer, typically some 10 nm of AlAs, between the LED structure and the substrate. A selective wet chemical etching process removes the sacrificial layers and separates the epitaxial layers from the GaAs substrate. Subsequently, the thin epitaxial film is transferred to a new wafer by van der Waals bonding [71]. A metal film on the carrier serves as a highly reflective back mirror for the LED. Surface roughening is achieved by depositing a monolayer film of randomly ordered polystyrene spheres on the LED wafer. Then the random pattern is transferred into the semiconductor surface by dry etching.

The principle of operation is to randomize the angles of totally reflected light at the top surface with the high reflectivity of a metallic back mirror on the bottom side (figure B1.1.21). Light that is not extracted through the surface is reflected back and forth between the back mirror and the rough surface. For high efficiency devices, the losses per round-trip through the LED structure have to be minimized by (i) reducing the thickness of all absorbing layers including the active layer and (ii) optimizing the reflectivity of the back mirror. Both can be achieved easily in the AlGaAs/InGaAs/GaAs material system. The band structure offers a sufficiently high carrier confinement to use only a few nanometres of active material for efficient light generation. Additionally the reflectivity of common metals like Au, Ag or Al is very high in the infrared regime.

The combination of epitaxial lift-off and surface roughening has been optimized by Windisch *et al* on near infrared (850 nm) LEDs [72, 73]. A schematic layout of the device is shown in figure B1.1.27*a*. The LED employs a mesa structure, with a selectively oxidized current aperture to prevent the generation of light under the metal contact. Current injection is facilitated via an annular top-side p-contact and a lateral n-contact around the mesa. In order to extract some of the laterally guided light, not only the surface on top of the mesa but also the area between mesa and bottom contact was roughened. Very high quantum efficiencies of up to 43 and 54% were achieved before and after encapsulation, respectively [74]. Current densities up to 1000 A cm^{-2} were applied. Similar devices at 650 nm emission wavelength resulted in 24% external quantum efficiency (31% after encapsulation) [75]. Due to relatively high forward voltages, the wall-plug efficiency of the red LEDs ranges between 10 and 15% at 1 mA of operation current.

B1.1.4.4 LED with tapered waveguide

A device that extracts solely the laterally guided modes inside the epitaxial structure is the LED with a radial tapered output coupler [76] shown in figure B1.1.27*b*. The circular device structure comprises a central top contact, a circular symmetric out-coupling taper with the shape of a shallow truncated cone and a ring-contact along the taper perimeter. Light extraction occurs through the bottom side, where the GaAs substrate has been removed by wet chemical etching.

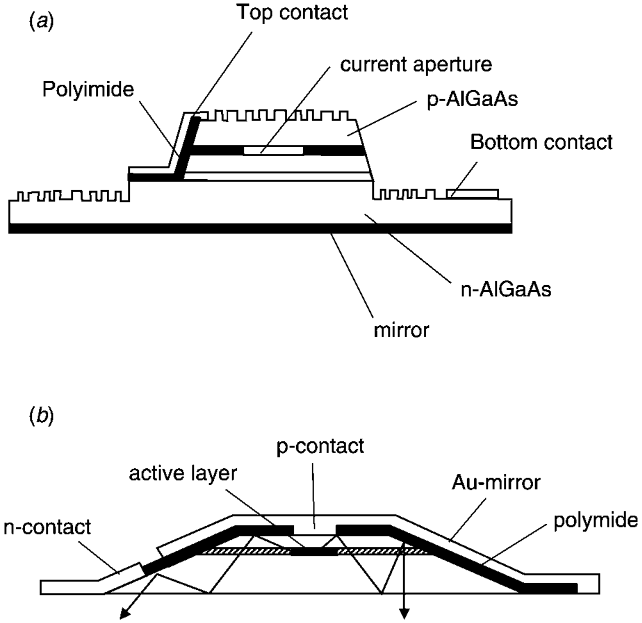


Figure B1.1.27. (*a*) Non-resonant LED with metallic back mirror and roughened surface. (*b*) LED with a radial tapered waveguide.

The principle of operation is to generate light in a small active area, defined by the geometry of the p-contact and to guide it radially towards the tapered area. The dimensions are chosen such that the light rays that hit the taper surface have a small azimuthal wave vector component which is an essential design parameter for efficient light extraction. Light that is reflected at the out-coupling surface is reflected again at the metal mirror on the taper and returns under a more favourable angle of incidence. The procedure is repeated until finally the angle of incidence falls within the escape cone. Absorption inside the waveguide layers which include unpumped active material and a reflectivity below 100% at the metal mirror on the taper limit the efficiency of this device.

With one and two InGaAs/GaAs compressively strained quantum wells as active material, Schmid *et al* demonstrated 45% quantum efficiency at 980 nm on encapsulated devices [77]. Maximum efficiency was reached at 2 mA injection current and a forward voltage around 2 V. This resulted in a wall-plug efficiency of 30 and 44%, before and after encapsulation, respectively. The first attempt to apply the same technology to 650 nm GaInP/AlGaInP devices resulted in an external efficiency of 13%.

B1.1.4.5 Resonant-cavity LEDs

The resonant-cavity LED or RCLED consists of a light-emitting active region between two Bragg mirrors which form an optical resonator. If the distance between the Bragg mirrors is set to a small multiple of half the optical wavelength, the cavity is in resonance with the emission and becomes transparent. In 1946, Purcell discovered that the spontaneous emission properties of an atom can be influenced by placing it inside a small (micro-) cavity [78]. The local strength of the electromagnetic modes and the density of states might then change the angular distribution as well as the emission rates. Schubert *et al* [79] proposed in 1992 to use a one-dimensional resonant cavity with a thin active layer inside as LED and to apply the cavity to manipulate the spontaneous emission.

The cavity determines the optical properties of the RCLED. It can be designed to direct more spontaneous emission into the escape cone than in the case of isotropic emission. This results in a higher intensity emitted in the surface normal direction as shown in figure B1.1.28. The different wavelengths within the emission linewidth fulfil the resonance conditions of the cavity at different angles of incidence at the top surface. Therefore, the RCLEDs spatial emission pattern contains every individual wavelength within the material linewidth in a specific angular direction. If the emission is spatially filtered, e.g. by coupling it into an optical fibre, also the range of emitted wavelengths is filtered, resulting in a narrow fibre coupled emission linewidth. However, when integrating over the entire half-sphere around the device, the measured emission linewidth is the same as that of the active material without the resonator.

A comprehensive introduction into the physics of RCLEDs can be found in the review of Benisty *et al* [80, 81]. Here, we will summarize only a few of the most important rules for the design of RCLEDs:

(1) The cavity order should be as low as possible. Even if the spacing between the Bragg mirrors is one half or one wavelength, the penetration depth of light into the Bragg layers increases the effective cavity length. Hence, the index contrast should be as high as possible.

(2) The quantum-well active layer should be placed in a maximum of the standing wave in the cavity.

(3) The emission of the active material should be tuned to a shorter wavelength than the resonance wavelength of the cavity. This compensates for a possible wavelength shift as the temperature rises and enhances the fraction of extracted light. The latter is due to the fact that the shorter wavelengths

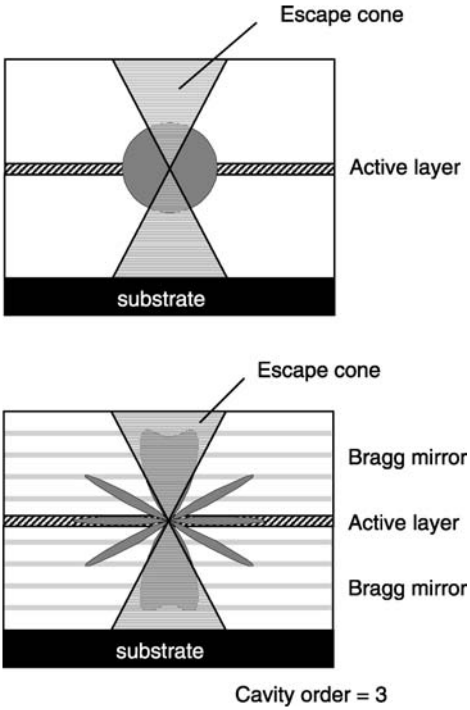


Figure B1.1.28. Spontaneous emission in a conventional LED (top) and in an RCLED (bottom). Because the spontaneous emission in the resonant cavity is no longer isotropic, the cavity can be designed to launch more emission into directions where the light is extracted (escape cone).

are resonant in off-axis directions with a much larger solid-angle element than the exactly tuned wavelength.

Contrary to conventional LEDs, many properties of RCLEDs such as the temperature dependence of intensity, the radiation pattern or the coupling efficiency into some kind of optical system can be designed. The key parameter here is the wavelength tuning of the cavity resonance, given by the thickness between the two mirrors and the emission wavelength. According to point 3, it is usually favourable to have the active material emitting at a few nanometres shorter than the on-axis cavity resonance wavelength. As the temperature is increased, the emission wavelength is shifted towards the resonance and increases the amount of light extracted in the surface normal direction. This is schematically shown in figure B1.1.29. Thus, for example the amount of fibre coupled light can increase with increasing temperature. At even higher temperatures, the emission will shift from the resonance to longer wavelengths and the surface normal intensity will decrease again. The wavelength tuning also affects the spatial radiation pattern of RCLEDs. If the peak intensity of the active layer emission is shifted to shorter wavelength than the cavity resonance, more light is extracted in off-normal directions.

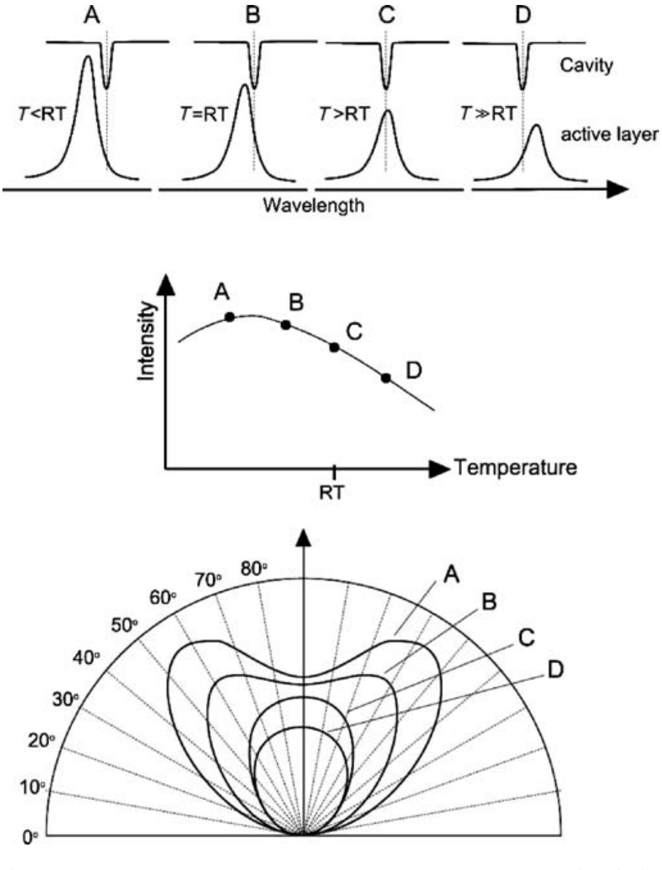


Figure B1.1.29. Effect of cavity tuning in RCLEDs. Top: emission wavelength of the active layer and cavity resonance at different temperatures. Middle: Emitted Intensity versus temperature. Bottom: Radiation pattern for the four temperatures A–D.

A two-dimensional-cross section of the resulting radiation pattern shows the typical double-lobe shape as shown in figure B1.1.29.

RCLEDs have been fabricated mostly for the infrared spectral range, often with the intention to use them as efficient light source in fibre communication systems. For this application, the RCLED has to compete with vertical cavity lasers, which are even more suitable for high-speed communication. However, a few groups continued to optimize RCLEDs and achieved remarkable results using, e.g. metal-mirrors or fully oxidized Al_xO_y/GaAs DBRs [82]. An external quantum efficiency above 20% was achieved by De Neve *et al* [83] with a bottom emitting 980 nm RCLED using an AlGaAs/GaAs-DBR and a metallic Ag-mirror. Also visible RCLEDs based on the AlGaInP/GaAs system have been demonstrated [16, 84, 85].

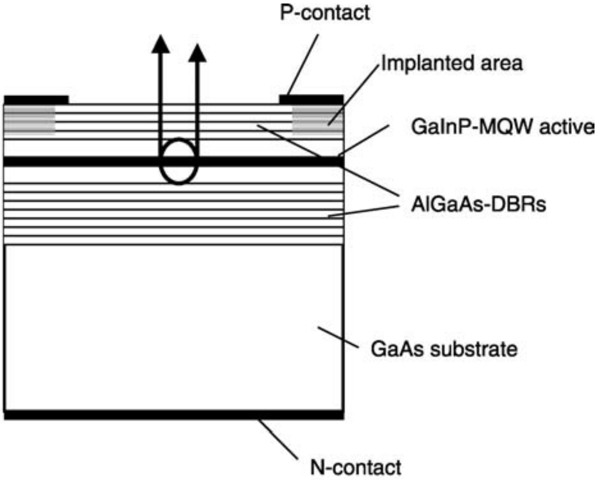
Only in the dark red around 650 nm, RCLEDs have made it to commercial products. At this wavelength there is a need for medium speed light sources for plastic fibre communication. RCLEDs are attractive because the directional output yields in a high fibre coupling efficiency and, due to the above-mentioned spatial filtering effect, in a narrow spectral linewidth in the fibre. Even with small devices and active diameters of less than 100 μm, levels of fibre coupled power well above 1 mW are achievable. If operated at some 10 mA of electrical current, the radiative carrier lifetime is reduced mainly by the high current density to rise- and fall-times of a few nanoseconds [86, 87]. Figure B1.1.30 shows the structure of a commercial 650 nm RCLED. It consists of two $Al_{0.5}Ga_{0.5}As/Al_{0.95}Ga_{0.5}As$ Bragg mirrors and an active layer of compressively strained GaInP quantum wells inside the $1-\lambda$-cavity [88]. For high-modulation speed, the material is processed to devices with an 80 μm light opening within an annular p-electrode (bottom of figure B1.1.29), but also conventional LED contact layouts are used. Large area devices achieve wall-plug efficiencies around 12% at 20 mA and 1.75 V forward voltage (figure B1.1.31). The high-speed devices are less efficient (9.5%), mainly due to the lack of photon recycling [89].

B1.1.4.6 Photonic bandgap approaches

Similar to the electronic bandgap, photonic bandgaps are defined as frequency bands over which all electromagnetic modes and spontaneous emission are suppressed [90–92]. They are a feature of periodically patterned materials with a strong dielectric constant contrast, so-called photonic crystals. The realization of photonic bandgap structures at optical frequencies requires the modulation of the dielectric constant on a nanometre scale, which is a challenging task for semiconductor processing technologies. In the context of optoelectronic device design, photonic bandgap structures are discussed as novel approaches for light confinement but their optical diffraction properties are also interesting.

Photonic bandgap structures can be employed for LEDs in different ways. One is to confine the photons in the device in two or three dimensions by using a periodic pattern in the semiconductor active material. This can be used to create waveguide structures with unique features on a very small scale but also to fabricate micro-cavities. The spontaneous emission in such a micro-cavity can be controlled if at least one dimension of the cavity is in the order of the emission wavelength. Ultimately, the vision is to realize single-mode LED emission or threshold-less micro-cavity lasing operation [92, 93]. In micro-cavities with metallic mirrors, two-dimensional photonic bandgap structures are used to create bandgaps for propagating surface waves or surface plasmon polaritons [94, 95]. Unlike three-dimensional photonic crystals, two-dimensional structures are easier to fabricate and therefore may be more interesting for practical applications.

Similar to grating output couplers on top-emitting in-plane lasers [96, 97], a two-dimensional photonic bandgap structure can be used to redirect laterally guided light inside an LED towards the surface normal direction. Photonic crystal light extractors have already been used in a number of different devices where the photonic bandgap structure was either etched into the top layer or all the way through the active LED structure. Erchak *et al* demonstrated a sixfold enhancement of PL-intensity at



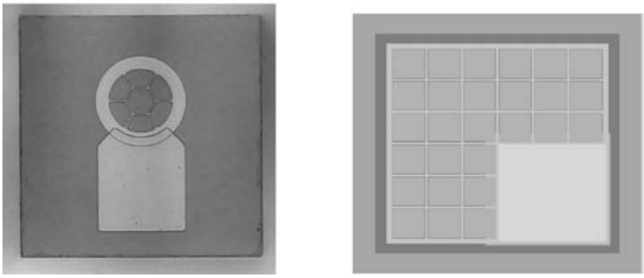


Figure B1.1.30. Top: Layout of a 650 nm RCLED with annular top contact and proton implanted current confinement. The distance between the AlGaAs-Bragg mirrors corresponds to one optical wavelength ($1 - \lambda$ cavity). Bottom: top view of two commercial RCLEDs. The chip dimensions are $(200 \times 250)\,\mu\text{m}^2$ and $(260 \times 260)\,\mu\text{m}^2$ for the left and right device. The smaller die has a light-emitting area with a diameter of 80 μm.

925 nm using a two-dimensional photonic bandgap in the upper cladding of an asymmetric LED structure [98]. Rattier [99] proposed a device where an unstructured LED-area for direct microcavity emission is surrounded by a deeply etched photonic bandgap for guided mode outcoupling. In both approaches, an oxidized Al_xO_y/GaAs Bragg reflector is used as bottom mirror, which offers a sufficiently high index contrast but is electrically isolating.

In order to combine an enhancement of the spontaneous emission rate with a mechanism for efficient light extraction, light can be generated in a thin film of active material and guided in a surface plasmon polariton mode [95]. These are guided optical modes that may exist at the interface between a dielectric and a metal and consist of an oscillating electromagnetic field coupled to an oscillating surface charge density. The rate of spontaneous emission can be increased by the coupling of the emitter to the enhanced electric fields associated with the guided modes (Purcell effect). A two-dimensional periodic photonic bandgap structure is used to create Bragg scattering and to couple the mode into useful radiation. A significant enhancement of PL extraction efficiency was demonstrated with optically pumped LEDs based on a thin-slab photonic crystal [100] and external quantum efficiencies of $>70\%$ were predicted with such a structure [94]. Most of the efficiency increase was associated with the improvement of Bragg extraction and to a minor increase in the Purcell enhancement.

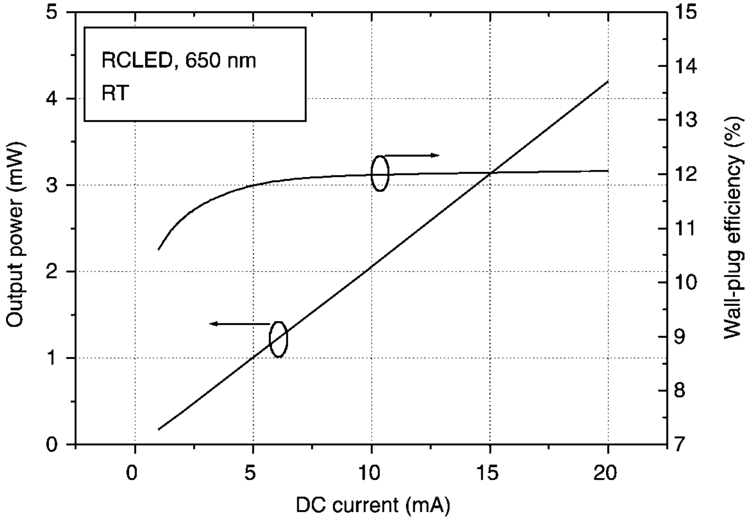


Figure B1.1.31. Wall-plug efficiency and optical output power of a 650 nm RCLEDs. The die size is 700 × 700 μm^2. At 20 mA, the device emits 4.2 mW.

The potential performance of LEDs using surface plasmon polariton modes is difficult to estimate. Apart from all the technical problems of realizing such devices, there are still a number of open questions to answer. Dissipative losses related to the propagation of light along a dielectric/metal interface will affect the performance and must be circumvented. Also it is not obvious if an effective extraction of the guided mode can be achieved with a structure that efficiently couples the emitter to the surface plasmon polarity mode [101].

B1.1.5 White LED

With the availability of wide-bandgap InGaN-based semiconductors, it became possible to produce not only efficient ultraviolet, blue and green diodes but also white emitting LEDs. Before that, the lack of white light has been a major barrier for LEDs in a wide range of applications. With LEDs covering the full colour spectrum including white, the ultimate goal of LED-based illumination and lighting became possible.

B1.1.5.1 White light

It is important to define what is meant by ‘white’ light, as it has to be compared with some kind of standard. The most fundamental model for white light is the solar spectrum shown in figure B1.1.32, but even then the sunlight changes with daytime and season of the year. It is usually therefore ‘idealized’ by using an equivalent black body radiation spectrum. According to Wien’s law, the peak maximum of the black-body spectrum is only a function of the temperature:

$$\lambda_{\text{peak}} = \frac{2880}{T}\ \mu\text{m K}. \tag{B1.1.40}$$

As the temperature increases from room temperature to thousands of degrees kelvin, the peak wavelength of the black body radiation moves from infrared to visible wavelengths. In the chromaticity

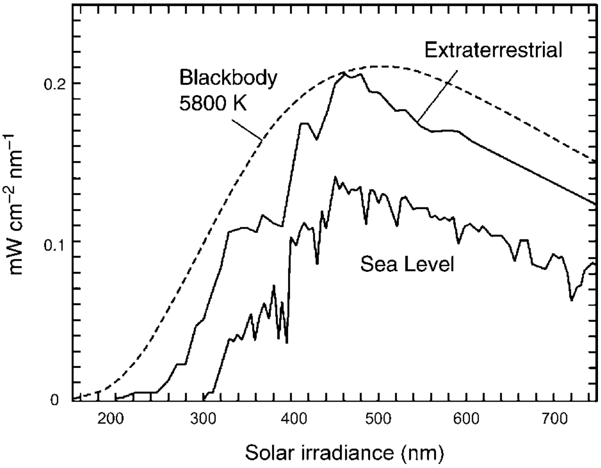


Figure B1.1.32. Spectrum of sunlight. The solid lines show the solar spectrum before and after transmitting the earth's atmosphere. The dashed line is the blackbody radiation at 5800 K.

diagram, this is represented by a line from the red towards the centre of the white colours (see appendix A). Consequently, assuming true black-body radiation, it is sufficient to use only one parameter, the *colour temperature*, to specify the spectral properties. The Commission Internationale de l'Eclairage (CIE) defined several standards for the so-called white, illuminants. The most commonly used standard for white light is the Illuminant C, describing overcast sunlight at a colour temperature around 6770 K.

Besides the colour temperature, the quality of white light can also be measured by the so-called *colour-rendering index* R_a. R_a describes the colour of objects illuminated with the light source as compared to the illumination with a reference light source. The CIE defined eight different sample objects, again using sunlight as reference. At first, eight special values R_a are deduced from these objects and then the average R_a is calculated as the arithmetic average [102]. By definition, sunlight has $R_a = 100$, incandescent bulbs achieve R_a up to 100, fluorescent lamps $R_a = 85$ and Na-vapour lamps $R_a = 20$. A selection of some application-specific requirements on R_a is given in table B1.1.5. The colour rendering index is a relative parameter and does not describe the real appearance of illuminated objects

Table B1.1.5. Colour rendering index required for different applications.

Application	R_a
Indoor/retail	90
Indoor office/home	80
Indoor work area	60
Outdoor pedestrian area	≥ 60
Outdoor general illumination	≤ 40

Source: ODIA workshop on LEDs for solid state lighting, October 26–27, 2000, Albuquerque, NM.

or the true colour of the light source. Therefore, even bluish or reddish light sources can achieve the ideal value of 100. Table B1.1.6 summarizes the values of R_a for a selection of conventional and LED-based white light sources. Note that the colour rendering index is very important for all illumination purposes, but is irrelevant for many other applications such as white signals or signs.

B1.1.5.2 Phosphor-converted white LEDs

The first commercial white LEDs introduced by Nichia, used the same principle as white-emitting high pressure mercury lamps, that of adding a 'colour-correcting' phosphor to the bulb to convert part of the primary emission into yellow. These LEDs mixed the residual blue emission from the LED chip with the complementary yellow luminescence from phosphor. The first phosphor-converted white LEDs were developed by the Fraunhofer Institute for Solid State Physics (IAF) in Germany. Blue emitting GaN LEDs were used in conjunction with organic colour convertors. The luminescent dyes, dissolved in the epoxy resin, absorbed the blue light and re-emitted luminescence at different colours [103]. Using the same principle of luminescence conversion, the blue GaN emission could also be used to fabricate green, yellow and red LEDs [104].

Today, white LED light is usually achieved by the use of an efficient inorganic phosphor and a single blue LED [105, 106]. The resulting emission spectrum, together with the human eye sensitivity is shown in figure B1.1.33. The most common phosphors are based on yttrium aluminium garnet (YAG) doped with one or more rare earth elements or rare earth oxides. Optically active rare earth elements such as neodymium (Nd), erbium (Er), cerium (Ce) or thorium (Th) are widely used in light sources like lasers, light amplifiers or fluorescent tubes. A very common and extensively studied phosphor for blue-yellow conversion is YAG:Ce ($Y_3Al_5O_{12}:Ce^{3+}(4f1)$) [107]. Within the packaging procedure, the YAG phosphor powder is suspended in the epoxy and deposited directly on or around the LED die [108]. Colour temperature and colour rendering index of such a single-chip white LED can be adjusted by the amount of phosphor suspended in the epoxy and the emission wavelength of the LED die. From a production point of view, the packaging step is critical for maintaining consistency in the colour characteristics, because subtle variations in concentration or spatial distribution of the phosphor coating changes the tint of white light between a more yellowish- or bluish-white.

Typically, the emission of such devices appears as a bluish or cold white due to the low level of absorption in the phosphor. Because the emission of the blue LED is directional, while the phosphor converted light radiates over a 2π solid angle, the appearance of white changes for an observer looking

Table B1.1.6. Colour rendering of different LED-based and conventional white light sources [111].

Technology	R_a
Blue LED + yellow phosphor	75–80
Blue LED + green + red phosphors	≥ 90
UV LED + blue + green + red phosphor	≥ 90
Red + blue + green LED (610/470/560 nm)	60
Blue + yellow LED (500/595 nm)	40
Halogen W-filament incandescent lamp	100
Fluorescent lamp	85
Na vapour lamp	20–65
Hg vapour lamp	35–55

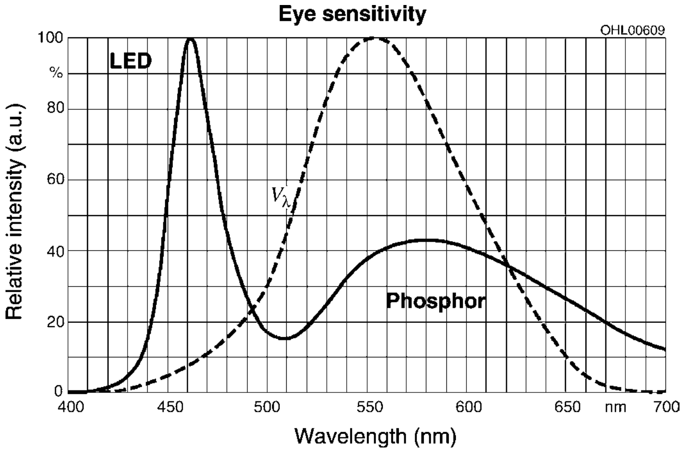


Figure B1.1.33. Spectral emission of a white LED based on phosphor conversion. Also shown is the human eye sensitivity with its maximum at 555 nm (green).

from the side. Compared to other approaches, phosphor converted LEDs have lower fabrication costs and offer a good control over quality and colour. Colour rendering with R_a around 75 is typically achieved at effective colour temperatures around 5000 K. The energy conversion efficiency of a single phosphor is below 60%, which is still rather low. Another drawback is the lifetime of the encapsulating materials which can degrade in the presence of the near-UV radiation tail from the blue LED die.

White light can also be generated by the combination of a near-UV (380–400 nm) LED and a two- or three-colour phosphor. The generation of white light by UV-pumped three-colour phosphors is widely used in fluorescent tubes and phosphors for the emission of the gas-discharge process at 254 nm are readily available. The overall efficiency of such an LED is given by the efficiency of the UV-LED, the degree of absorption in the phosphors and the efficiency of the wavelength-conversion process. Despite the challenging material constraints in the AlGaN/InGaN system [109], UV LEDs with remarkably high quantum efficiencies of up to 32% at 390 nm and output power levels of 21 mW at 20 mA have been reported [110]. Although less efficient than these record values, UV InGaN-LEDs with a few mW of output power are already commercially available. The UV absorption in the phosphors is still an issue for optimization and much effort is dedicated to either developing LEDs that match the absorption of existing phosphors or developing appropriate phosphors absorbing at wavelengths where the LED is most efficient. However, even with 100% UV-absorption, a substantial amount of energy is lost in the actual wavelength conversion processes which markedly limits the overall luminous performance of such devices.

A potential advantage of the UV-pumped phosphor-based LED is that the visible emission is solely originated by phosphorescence with a broad spectral output. The superposition of three broad lines at red, green and blue (e.g. 610, 560 and 470 nm) results in white light with excellent CRI values (>90). Also, the white point in the CIE chromaticity diagram is independent of the LED characteristics. Thus, in terms of colour rendering and colour stability the UV pumped phosphor-based LED is the most attractive source for white lighting.

B1.1.5.3 Multi-chip white

By definition, the addition of two complementary colours produces white. Hence, white light can also be generated by mixing the emission of two different LEDs with complementary colours. In the chromaticity diagram, the positions of complementary colours lie on opposite sides of the achromatic point with $x = y = 0.3333$. All white points along this line are accessible by adjusting the power ratio of the two monochromatic colours at the perimeter of the chromaticity diagram. An example for white light generated by a mixture of blue (490 nm) and amber (590 nm) is shown in figure B1.1.34. LED-based white light sources using two discrete colours are very efficient in terms of luminous performance ($\mathrm{lm\,W^{-1}}$) and costs/lumen. The main problem, however, is the very limited colour rendering which excludes this solution for general lighting applications. Despite this drawback, two-chip white LEDs are commercially available and used, e.g. as map lights in the interior of automobiles. The luminous output of such lamps can be increased by using several LEDs of each colour within the same package.

White with a better spectral distribution is generated by the mixture of three discrete colours of three different LED chips. The three-chip white LED has the advantage of providing no inherent loss mechanisms due to wavelength conversion to the emission from the primary sources. In the CIE chromaticity diagram, the three monochromatic colours span a triangle in which all mixed colours including white are accessible by the adjustment of the power ratios of the corner point colours. Figure B1.1.38 shows an arrangement of three LEDs (blue, green and red) in a single package. With individually addressable dices, such a multi-chip full colour LED lamp is capable of producing various mixed colours, including a range of whites. As a white light source, the multi-chip LED achieves relatively good colour rendering, colour temperatures between 3000 and 7000 K, and high efficiencies in terms of luminous performance [112]. Unlike single-chip white LED, the colour characteristics of multi-chip white LEDs can be altered after the packaging step. Used only as a source of white light, the multi-chip LED might be too expensive, but it is ideally suited for applications where variable-colour pixels are required. Commercial multi-chip LEDs are offered with three or even five dices of discrete colour in one lamp.

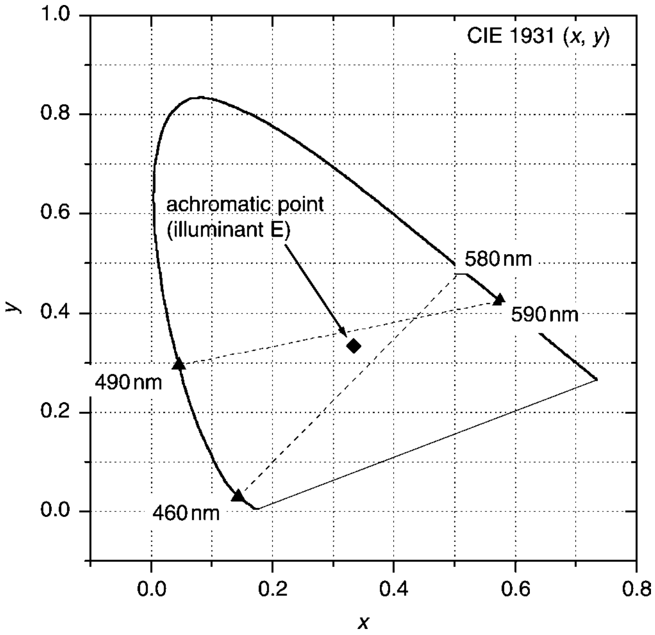


Figure B1.1.34. Examples for complementary colours in the CIE chromaticity diagram. The line between 490 and 590 nm is used for two-chip white LEDs. The 460–580 nm line is an example for white from phosphor-converted blue.

B1.1.6 Applications

B1.1.6.1 Packaging

The packaging technology is becoming increasingly important for the performance of LEDs in many current and future applications. Some of the older, conventional packages today are inadequate for the rapidly improving high brightness AlGaInP or InGaN dices. Novel packages must consider better optical, electrical and thermal performance. The demand for high reliability puts stringent requirements on the chemical and thermal stability of the packaging, die attach and encapsulating materials as well as the selected processes. Devices with light converting phosphors such as white LEDs have to take additional care of the efficiency and stability of the phosphor materials. The best monochromatic LEDs today achieve $>100\,\mathrm{lm\,W^{-1}}$ and the goal to achieve this and better efficiencies within the entire visible spectrum and, most importantly, in white LEDs will depend not only on the dices but also on the package used for assembly.

A simple radial LED package is shown in figure B1.1.35. It was originally designed for low current indoor applications and has a thermal resistance of max. $280\,\mathrm{K\,W^{-1}}$ (R_{thjs}, thermal resistance at junction solder point), limiting the electrical input power to some hundred mW. The die has typically a lateral dimension of 200–300 μm. It is attached to the metal lead frame with epoxy-based conductive glue, so that the lead frame can act as electrical contact to the outside. This metal is shaped in a way that it can act as a mirror cup as well as heat sink for the die. A separate metal pin is connected to the bond-pad on the chip and acts as second electrode. Chip and lead-frame are encapsulated by epoxy which has the form of a dome in order to achieve a certain radiation characteristics. The epoxy almost doubles the extraction efficiency due to its favourable refractive index around 1.5 and the non-planar epoxy–air interface. Standard diameters of the epoxy dome are 3 and 5 mm and the package is also named simply '5 or 3 mm-LED package'. About two-thirds of all high brightness LEDs are shipped in radial housings.

Packages for LEDs, like for most electronic components, can be divided into two categories: through hole and surface mount. Through hole components like the radial package are loaded to a PC-board from one side and soldered from the other. Surface mount devices (SMDs) are loaded and soldered on the same side. This has several benefits for industrial production such as faster placing in automatic machines, smaller size, less parasitic effects and lower costs. In particular, in applications where space is limited such as in mobile phones, the surface mount technology (SMT) is superior. Figure B1.1.36

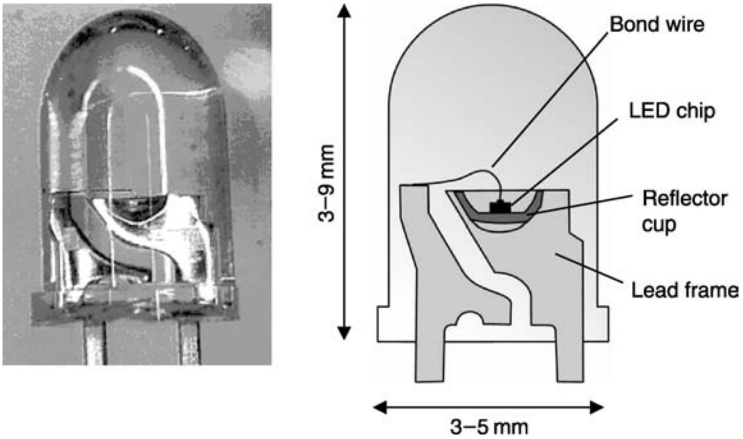


Figure B1.1.35. Radial LED lamp.

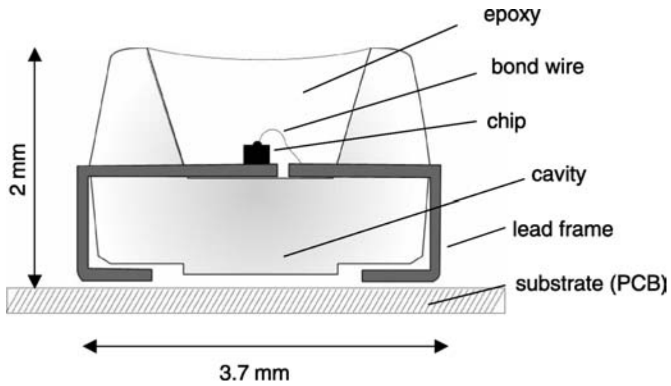


Figure B1.1.36. Schematic drawing of a surface mount LED package (TOPLED).

shows a standard SMT-package for LEDs with the lateral dimensions of $3.4 \times 3.0\,\mathrm{mm}^2$ and a height of 2.1 mm. The die is attached to the lead-frame with the pre-moulded housing of plastic (A-model). A bond wire connects the die's top electrode to the other part of the lead-frame. Finally, the housing is filled with encapsulating resin and, if necessary, one or more phosphor converters. The flat epoxy–air interface of SMT-packages results in a 10% lower extraction efficiency compared to radial packages with epoxy domes. However, for applications with the need for narrower emission profiles or more directionality, a transparent lens can be integrated (see figure B1.1.37).

Depending on the application, the leads are folded either outwards or inwards under the housing. Some packages use specially bent leads so that the LED can disappear in a hole in the PC-board while still being soldered from the top. SMT packages can be made very small. The smallest devices are 0.5–1 mm wide and high, which is just a little more than the dimensions of the die. The thermal resistance of SMT packages ranges from 300 to $500\,\mathrm{K\,W^{-1}}$. This limits the maximum applicable current to 100–150 mA. SMT packages can also house several chips, e.g. for the generation of white light or as multiple-colour LEDs. An example for a multi-chip SMT package is shown in figure B1.1.38.

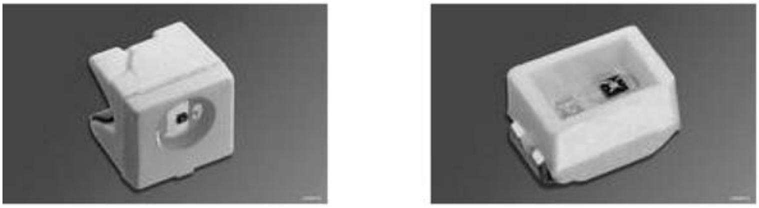
High-flux LEDs are designed for operation currents of 1 A or even more. For this current range, the package has to be capable of dissipating thermal powers of 1–2 W. High power packages, like the one shown in figure B1.1.39, include a deep reflector and heat sinking metal base. The package can also include a lens-shaped epoxy dome to optimize the radiation pattern.

B1.1.6.2 Applications

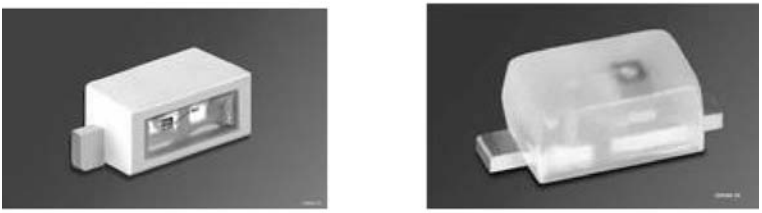
For a long time, LEDs were low cost, low brightness devices suitable for single colour, low power applications, e.g. to illuminate switches, indicators, small signs or to transmit information. With the availability of high brightness yellow and red AlGaInP LEDs and the development of blue and green InGaN-LEDs, the full colour spectrum could be covered and a whole range of new applications opened up. A good example is signs and displays, where originally LEDs were only applied for monochromatic indoor numerical or text displays. Today, LED-based single or full colour outdoor signs can be found as variable message signs on motorways or large area video panels for entertainment arenas and sports stadiums. One of the largest high resolution outdoor panels is the outer skin of the NASDAQ-building at New York's Time Square, which consumes 18 million LEDs. The LED sign market accounts for about one-third of the total market for high brightness LEDs (2000: $1.22 billion).

LCD backlighting—The rising market for cellular phones has also opened up new opportunities for LEDs. Here, LEDs are used to illuminate both the LCD display as well as the keypad. In 2000 about 400

TOPLED TOPLED with lens



SIDELED Mini TOPLED



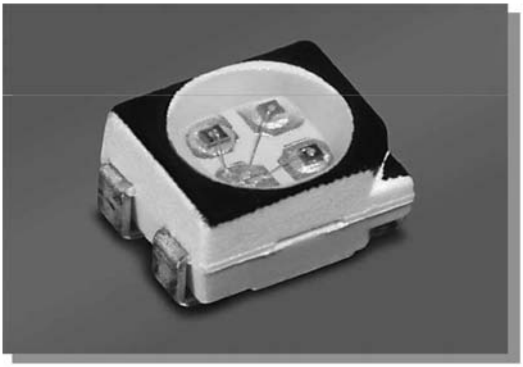
Micro SIDELED SmartLED

Figure B1.1.37. Examples for SMT packages. Note the small size of, e.g. the SmartLED: 0.8 × 1.2 and 0.6 mm height.

million cellular phones were sold worldwide with an average number of approximately ten LEDs per unit. This application is still dominated by standard-type low brightness GaP yellow-green LEDs, but the use of high-brightness devices, in particular white, is growing rapidly. The arguments for LEDs here is mainly their small size but also the low power consumption, which helps to prolong the battery lifetime. White LEDs are also increasingly used to illuminate other types of LCD displays such as digital camera displays, handheld computers or, maybe in a longer perspective, even large area LCD screens.

Traffic lights—Traffic lights are a typical example for an application, where LEDs as efficient monochromatic light sources can favourably compete with conventional solutions. A tungsten bulb with a red filter provides about 4 lm W^{-1} compared to more that 50 lm W^{-1} of a high-brightness AlGaInP-diode. The motivation to use LEDs is mainly driven by energy savings and reduced requirements for replacement. After the initial hurdles of standards and regulations were taken, high brightness AlGaInP-(red and yellow) and InGaN-LEDs (green) have an impressive degree of market penetration, especially in the US. In 2000 about 14% of all red and 6% of all green traffic signals in the US had been converted to high-brightness LEDs [113].

LEDs for automotive applications—Another application which increasingly consumes high numbers of LEDs is the interior and exterior illumination of automobiles. Inside a car, LEDs are used not onlly to backlight the dash panel, push buttons and indicator lamps but also LCD-displays, e.g. for the navigation system. Primarily, LEDs are used for dashboard illumination because their lifetime is



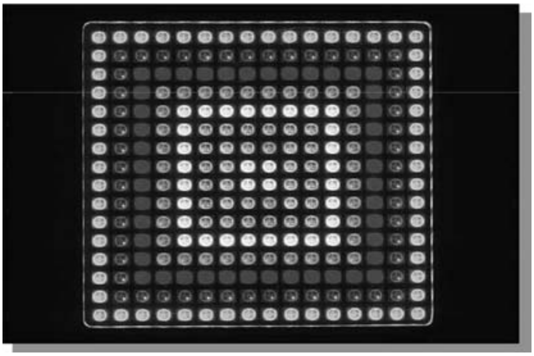


Figure B1.1.38. Three-chip LED in an SMT-package. Below: arrangement of MultiLEDs for a full-colour display.

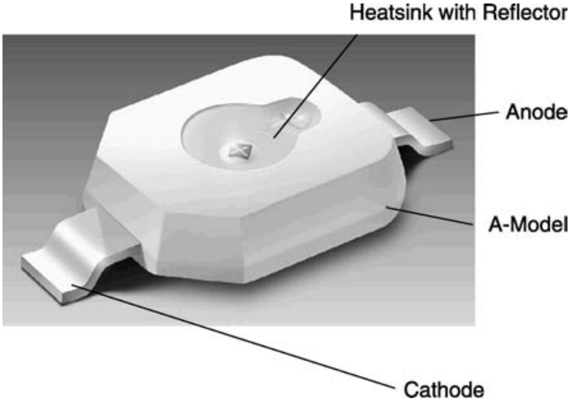

Heatsink with Reflector
Anode
A-Model
Cathode

Figure B1.1.39. High flux package. The package can take dices up to 1 × 1 mm^2 area and dissipate more than 1 W of excessive heat.

significantly longer than that of conventional incandescent bulbs. Despite the higher price for LEDs, cost savings can be achieved, because no replacement has to be considered, but also because surface mount LED packages allow for cost efficient automated dashboard assembly. In total, the number of LEDs inside a car can easily sum up to several hundred pieces per car.

The first application in the exterior was the centre high mounted stop light (CHMSL), which is still a steadily increasing market for AlGaInP red LEDs. While the CHMSL was for many years the only exterior LED application for automobiles, LEDs have now also started to penetrate other areas such as stop- or tail-lights and indicators. The benefits of LEDs here are their ruggedness, small size, low power consumption and the high reliability. They also offer a much faster turn-on time compared to incandescent bulbs, which is an important safety aspect for stop-lights and CHMSLs. Sometimes, LEDs are clustered in modules of 70 or 80 red and amber LEDs that combine the functions of stop-, tail- and indicator lights. So far only a few car manufacturers have started to equip some high-end car models with LED-based exterior functions, but it is expected that the number of LEDs in the exterior of automobiles and trucks will grow rapidly over the next years.

General lighting—The ultimate goal for the future of white LEDs is, of course, general lighting. This market is valued to approximately $12 billion and represents a huge potential market for LEDs. In order to make this happen, LEDs do have to face a number of issues that are relevant to the lighting industry. The efficiency of LEDs as well as the light-output generated per device or module has to be raised to the level of conventional lighting systems. In terms of efficiency, the target for white LEDs is somewhere between 100 and 150 $\mathrm{lm\,W^{-1}}$, the total generated flux should come up at least to levels of incandescent bulbs (1700 lm for a 100 W bulb). For indoor lighting, the colour rendering index R_a describing the appearance of colours when illuminated with LED light, should be higher than 80. The colour temperature of LED-white should at least replicate the values of fluorescent or incandescent (2850 K) lamps. Colour variations in different directions of radiation or variations from device to device have to stay within the acceptable ranges of conventional sources. The colour stability of phosphor-converted white LEDs is mainly an issue for the packaging process, whereas multi-chip LEDs require very reproducible intensity ratios or an active adjustment of the colour point. A long device lifetime is important to reduce the overall operating costs including replacement and maintenance. The crucial part for high reliability is the package including encapsulating and converter materials. Even if the diode chip alone has a lifetime of many ten thousand hours, the encapsulating materials tend to degrade faster in the presence of UV-photons. Finally, it is of paramount importance to reduce the cost for white LED light to competitive levels.

Acknowledgment

I would like to thank all colleagues at Osram Opto Semiconductors who contributed to this chapter. In particular, I thank the AlGaInP R&D team: Norbert Linder, Ralph Wirth, Walter Wegleiter, Reiner Windisch, Peter Stauss, Christian Karnutsch, Wolfgang Schmid, Ines Pietzonka, Simone Thaler, Gertraud Huber, Monika Mändl, Kornelia Kruger and Monika Kuttenberger.

References

[1] Round H J 1907 A note on carborundum *Electrical World* **49** 309
[2] Pankove J I, Maruska H P and Berkeyheiser J E 1970 Optical absorption of GaN *Appl. Phys. Lett.* **17** 197–199
[3] Roosbroeck W V and Shockley W 1954 Photon-radiative recombination of electrons and holes in Ge *Phys. Rev. B.* **94** 1558–1560
[4] Olshanksy R, Su C, Manning J and Powazi W 1984 Measurement of radiative and nonradiative recombination rates in InGaAsP and AlGaAs light sources *J. Quantum Electron.* **20** 838–854

[5] Henry C H, Logan R A and Merrit F R 1978 The effect of surface recombination on current in AlGaAs heterojunctions *J. Appl. Phys.* **49** 3530–3542

[6] Boroditsky M, Gontijo I, Jackson M, Vrijen R and Yablonovitch E 2000 Surface recombination measurements on III–V candidate materials for nanostructure light-emitting diodes *J. Appl. Phys.* **87** 3497–3504

[7] Saul R H, Lee T P and Burrus C A 1985 *Light-Emitting-Diode Device Design in Semiconductors and Semimetals* vol 22 (Bell Telephone Laboratories Inc) (San Diego: Academic)

[8] Born M. W E 1959 *Principles of Optics* (Cambridge: Cambrige University Press)

[9] Carr W and Pittman G 1963 One-watt GaAs p–n junction infrared source *Appl. Phys. Lett.* **3** 173–175

[10] Franklin A and Newman R 1964 Shaped electroluminescent GaAs diodes *J. Appl. Phys.* **35** 1153–1155

[11] Osram enhances brightness of blue InGaN LEDs *Compound Semicond.* 2001 vol 7, p 7

[12] Gardner N F, Chui H C, Chen E I, Krames M R, Huang J W, Kish F A, Stockman S A, Kocot C P, Tan T S and Moll N 1999 1.4x efficiency improvement in transparent-substrate $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ light-emitting diodes with thin (<2000 Å) active regions *Appl. Phys. Lett.* **74** 2230–2232

[13] Linder N, Kugler S, Stauss P, Streubel K P, Wirth R and Zull H 2001 High-brightness AlGaInP light-emitting diodes using surface texturing *Proc. SPIE* **4278** 19–25

[14] Krames M R *et al* 1999 High-power truncated-inverted-pyramid $(Al_xGa_{1-x})_{0.5}In_{0.5}P$/GaP light-emitting diode exhibiting $>50\%$ external quantum efficiency *Appl. Phys. Lett.* **75** 2365–2367

[15] Härle V, Hahn B, Lugauer H, Brüderl G, Eisert D, Strauss U, Lee A and Hiller N 2000 Developments in InGaN on SiC LEDs and lasers at Osram *Compound Semicond.* **6** 81–85

[16] Streubel K, Helin U, Oskarsson V, Bäcklin E and Johansson A 1998 High brightness visible (660 nm) resonant cavity light emitting diode *Photon. Technol. Lett.* **10** 1685–1687

[17] Nishizawa J, Koike M and Jin C C 1983 Efficiency of GaAlAs heterostructure red light-emitting diode *J. Appl. Phys.* **54** 2807–2812

[18] Ishiguro H, Sawa K, Nagao S, Yamanaka H and Koike S 1983 High efficient GaAlAs light-emitting diodes of 660 nm with a double heterostructure on a EaAlAs substrate *Appl. Phys. Lett.* **43** 1034–1036

[19] Shockley W 1950 *Electrons and Holes in Semiconductors* (New Jersey: Princeton)

[20] Sze S M 1981 *Physics of Semiconductor Devices* (New York: Wiley)

[21] Craford M G, Shaw R W, Herzog A H and Groves W O 1972 Radiative recombination mechanisms in GaAsP diodes with and without nitrogen doping *J. Appl. Phys.* **43** 4075–4083

[22] Campel J C, Holonyak N, Craford M G and Keune L D 1974 Band structure enhancement and optimization of radiative recombination in $GaAs_{1-x}P_x$:N (and $In_{1-x}Ga_x$ P:N) *J. Appl. Phys.* **45** 4543–4553

[23] Casey H C and Panish M B 1978 *Heterostructure Lasers: Part A. Fundamental Principles* (New York: Academic)

[24] Watanabe M O, Yoshida J, Mashita M, Nakashini T and Hojo A 1985 Band discontinuity for GaAs/AlGaAs heterojunction determined by C–V profiling technique *J. Appl. Phys.* **57** 5340–5345

[25] Steranka FM 1997 AlGaAs red light-emitting diodes *Semiconductors and Semimetals, High Brightness Light Emitting Diodes* vol 48, eds G B Stringfellow and M G Craford pp 65–95

[26] Meney A, Prins A, Philips A, Sly J, O'Reilly E, Dunstan D, Adams A and Valster A 1995 Determination of the band structure of disordered AlGaInP and its influence on visible laser characteristics *J. Quantum Electron.* **1** 697–706

[27] Brennan K and Chiang P K 1992 Calculated electron and hole steady-state drift velocities in lattice matched GaInP and AlGaInP *J. Appl. Phys.* **71** 1055

[28] Kish F and Fletcher R 1997 AlGaInP light-emitting diodes *Semiconductors and Semimetals, High Brightness Light Emitting Diodes* vol 48, eds G B Stringfellow and M G Craford pp 149–220

[29] Morrison A P, Lambkin J D, Poel C J and Valster A 2000 Electron transport across bulk (AlGa)InP barriers determined from the *I–V* characteristics of n–i–n diodes measured between 60 and 310 K *J. Quantum Electron.* **36** 1293–1298

[30] Streubel K, Linder N, Wirth R and Jaeger A 2002 High brightness AlGaInP LEDs *J. Sel. Top. Quantum Electron.* **8** 321–332

[31] Chang S J, Chang C S, Su Y K, Chang P T, Wu Y R, Huang K H and Chen T P 1997 AlGaInP yellow-green light emitting diodes with a tensile strain barrier cladding layer *Photon. Technol. Lett.* **9** 1199–1201

[32] Zunger A and Mahajan S 1995 *Handbook of Semiconductors* (Amsterdam: Elsevier)

[33] Su L C, Ho I H, Kobayashi N and Stringfellow G B 1994 Order/disorder heterostructure in $Ga_{0.5}In_{0.5}P$ with $\Delta E_g = 160$ meV *J. Cryst. Growth* **145** 140–146

[34] Ernst P, Geng C, Scholz F, Schweizer H, Zhang Y and Mascarenhas A 1995 Band-gap reduction and valence-band splitting of ordered GaInP *Appl. Phys. Lett.* **67** 2347–2349

[35] Ernst P, Geng C, Scholz F and Schweizer H 1996 Ordering in GaInP studied by optical spectrometry *Phys. Stat. Sol. (b)* **193** 213

[36] Nasi L, Salviati G, Mazzer M and Zanotti-Fregonara C 1996 Influence of surface morphology on ordered GaInP structures *Appl. Phys. Lett.* **68** 3263–3265

[37] Suzuki T, Gomyo A, Hino I, Kobayashi K, Kawata S and Iijima S 1988 P-type doping effects on band-gap energy for GaInP grown by metalorganic vapor phase epitaxy *Japan. J. Appl. Phys. Lett.* **27** L1549–L1552

[38] Vanderwater D A, Tan I H, Höfler G E, Defevere D C and Kish F A 1997 High-brightness AlGaInP light emitting diodes *Proc. IEEE* **85** 1752–1764

[39] Koide Y, Itoh H, Khan M R H, Hiramatsu K, Sawaki N and Akasaki I 1987 Energy band-gap bowing parameter in an $Al_xGa_{1-x}N$ alloy *J. Appl. Phys.* **61** 4540–4543

[40] Ho I H and Stringfellow G B 1996 Solid phase immiscibility in GaInN *Appl. Phys. Lett.* **69** 2701–2703

[41] Wright A F and Nelson J S 1995 Bowing parameters for zinc-blende $Al_{1-x}Ga_xN$ and $Ga_{1-x}In_xN$ *Appl. Phys. Lett.* **66** 3051–3053

[42] Lester S D, Ponce F A, Craford M G and Steigerwald D A 1995 High dislocation densities in high efficiency GaN-based light-emitting diodes *Appl. Phys. Lett.* **66** 1249–1251

[43] Härle V, Hahn B, Lugauer H, Bader S, Brüderl G, Baur J, Eisert D, Strauss U, Zehnder U, Lell A and Hiller N 2000 GaN-based LEDs and lasers on SiC *Phys. Stat. Sol. (a)* **180** 5–13

[44] Stath N, Härle V and Wagner J 2001 The status and future development of innovative optoelectronic devices based on III-nitrides on SiC and on III-antimonides *Mater. Sci. Eng.* **B80** 224–231

[45] Yoshida S, Misawa S and Gonda S 1983 Improvements on the electrical and luminescent properties of reactive molecular beam epitaxially grown GaN films by using AIN-coated sapphire substrates *Appl. Phys. Lett.* **42** 427–429

[46] Amano H, Sawaki N, Akasaki I and Toyoda Y 1986 Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AIN buffer layer *Appl. Phys. Lett.* **48** 353–355

[47] Nakamura S 1991 GaN growth using GaN buffer layer *Japan. J. Appl. Phys. Lett.* **30** 1705–1707

[48] Nakamura S 1997 Group III–V nitride-based ultraviolet LEDs and laser diodes *Semiconductors and Semimetals, High Brightness Light Emitting Diodes*, eds G B Stringfellow and M G Craford pp 391–441

[49] Strite S, Lin M E and Morkoc H 1993 Progress and prospects for GaN and the III–V nitride semiconductors *Thin Solid Films* **231** 197–210

[50] Maruska H P and Tietjen J J 1969 The preparation and properties of vapor-deposited single-crystal-line GaN *Appl. Phys. Lett.* **15** 327–329

[51] Amano H, Kitoh M, Hiramatsu K and Akasaki I 1990 Growth and luminescence properties of Mg-doped GaN prepared by MOVPE *J. Electrochem. Soc.* **137** 1639–1641

[52] Nakamura S, Iwasa N, Senoh M and Mukai T 1992 Hole compensation mechanism of P-type GaN films *Japan. J. Appl. Phys. Lett.* **31** 1258–1266

[53] Nakamura S, Mukai T and Senoh M 1994 Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes *Appl. Phys. Lett.* **64** 1687–1689

[54] Nakamura S, Mukai T and Senoh M 1994 High-brightness InGaN/AlGaN double-heterostructure blue-green-light-emitting diodes *J. Appl. Phys.* **76** 8189–8191

[55] Akasaki I, Amano H, Koide Y, Hiramatsu K and Sawaki N 1995 Crystal growth of column III nitrides and their applications to short wavelength light emitters *J. Cryst. Growth* **146** 455–461

[56] Nakamura S 1995 High-brightness InGaN blue, green and yellow light-emitting diodes with quantum well structures *Japan. J. Appl. Phys. Lett.* **34** L797–L799

[57] Björk G, Yamamoto Y and Heitman H 1995 *Confined Electrons and Photons* (New York: Plenum)

[58] Joannoplolus J D, Meade R D and Winn J N 1995 *Photonic Crystals* (Princeton: Princeton University Press)

[59] Kish F A *et al* 1994 Very high-efficiency semiconductor wafer-bonded transparent-substrate $(Al_xGa_{1-x})_{0.5}In_{0.5}P$/GaP light-emitting diodes *Appl. Phys. Lett.* **64** 2839–2841

[60] Kish F A, Vanderwater D A, Peanasky M J, Ludowse M J, Hummel S G and Rosner S J 1995 Low-resistance Ohmic conduction across compound semiconductor wafer-bonded interfaces *Appl. Phys. Lett.* **67** 2060–2062

[61] Höfler G E, Vanderwater D, DeFevere D C, Kish F A, Carnras M, Steranka F and Tan I H 1996 Wafer bonding of 50 mm diameter GaP to AlGalnP-GaP light-emitting diode wafers *Appl. Phys. Lett.* **69** 803–805

[62] Carr W N 1966 Photometric figures of merit for semiconductor luminescent sources operating in spontaneous mode *Infrared Phys.* **6** 1–19

[63] Illek S, Jacob U, Plößl A, Stauss P, Streubel K, Wegleiter W and Wirth R 2002 Buried micro-reflectors boost performance of AlGaInP LEDs *Compound Semicond.* **8** 39–42

[64] Horng R H, Wuu D S, Wei S C, Tseng C Y, Huang M F, Chang K H, Liu P H and Lin K C 1999 AlGaInP light-emitting diodes with mirror substrates fabricated by wafer bonding *Appl. Phys. Lett.* **75** 3054–3056

[65] Horng R H, Wuu D S, Wei S C, Huang M F, Chang K H, Liu P H and Lin K C 1999 AlGAInP/AuBe/glass light-emitting diodes fabricated by wafer bonding technology *Appl. Phys. Lett.* **75** 154–156

[66] Härle V H 2003 Light extraction technologies for high efficiency GaInN-LED devices *SPIE Photonics West'03* (San Jose, USA, Jan. 2003)

[67] Schnitzer I, Yablonovitch E, Caneau C and Gmitter T J 1992 Ultra high spontaneous emission quantum efficiency, 99.7% internally and 72% externally, from AlGaAs/GaAs/AlGaAs double heterostructures *Appl. Phys. Lett.* **62** 131–133

[68] Schnitzer I and Yablonovitch E 1993 30% external quantum efficiency from surface textured, thin-film light-emitting diodes *Appl. Phys. Lett.* **63** 2174–2176

[69] Yablonovitch E, Gmitter T, Harbison J P and Bhat R 1987 Extreme selectivity in the lift-off of epitaxial GaAs films *Appl. Phys. Lett.* **51** 2222–2224

[70] Deckman H W and Dunsmuir J H 1982 Natural lithography *Appl. Phys. Lett.* **41** 377–379

[71] Yablonovitch E, Hwang D M, Gmitter T J, Florenz L T and Harbison J P 1990 van der Waals bonding of GaAs epitaxial liftoff films onto arbitrary substrates *Appl. Phys. Lett.* **56** 2419–2421

[72] Windisch R, Kuijk M, Dutta B, Knobloch A, Kiesel P, Döhler G H, Borghs G and Heremans P 2000 Non-resonant cavity light-emitting diodes *SPIE Proc. Photonics West* 1–7

[73] Windisch R, Dutta B, Kuijk M, Knobloch A, Meinlschmidt S, Schobert S, Kiesel P, Borghs G, Döhler G H and Heremans P 2000 40% efficient thin-film surface-textured light-emitting diodes by optimization of natural lithography *IEEE Trans. Electron. Devices* **47** 1492–1498

[74] Windisch R, Rooman C, Meinlschmidt S, Kiesel P, Zipperer D, Döhler G H, Dutta B, Kuijk M, Borghs G and Heremans P 2001 Impact of texture-enhanced transmission on high-efficiency surface-textured light-emitting diodes *Appl. Phys. Lett.* **79** 1–3

[75] Rooman C, Windisch R, D'Hondt M, Dutta B, Modak P, Mijlemans P, Borghs G, Vounckx R, Moerman I, Kuijk M and Heremans P 2001 High-efficiency thin-film light emitting diodes at 650 nm *Electron. Lett.* **37** 852–853

[76] Schmid W *et al* 1999 Infrared light-emitting diodes with lateral outcoupling taper for high extraction efficiency *SPIE Photonics West'99* (San Jose, USA, Jan. 1999).

[77] Schmid W, Eberhard F, Jäger R, King R, Miller M, Joos J and Ebeling K J 2000 45% Quantum efficiency light-emitting diodes with radial outcoupling taper *SPIE Proc. Photonics West* **3938** 90–97

[78] Purcell E M 1946 Spontaneous emission probabilities at radio frequencies *Phys. Rev. Lett.* **69** 681

[79] Schubert E F, Wang Y H, Cho A Y, Tu L W and Zydzik G J 1992 Resonant cavity light-emitting diode *Appl. Phys. Lett.* **60** 921–923

[80] Benisty H, Neve H D and Weisbuch C 1998 Impact of planar microcavity effects on light extraction—Part I: Basic concepts and analytic trends *J. Quantum Electron.* **34** 1612–1631

[81] Benisty H, Neve H D and Weisbuch C 1998 Impact of planar microcavity effects on light extraction—Part II: Selected exact simulations and role of photon recycling *J. Quantum Electron.* **34** 1632–1643

[82] Huffaker D L, Lin C C, Shin J and Deppe D G 1995 Resonant cavity light emitting diode with an Al_xO_y/GaAs reflector *Appl. Phys. Lett.* **66** 3096–3098

[83] De Neve H, Blondelle J, Baets R, Demeester P, Daele P V and Borghs G 1996 Resonant cavity LEDs *Microcavities Photonic Bandgaps: Physics Appl.* **324** 333–342

[84] Jalonen M, Toivonen M, Köngäs J, Savolainen P, Salokatve A and Pessa M 1997 Oxide-confined AlGaInP/AlGaAs visible resonant cavity light-emitting diodes grown by solid source molecular beam epitaxy *LEOS 10th Annual Meeting 97* (Nov. 1997) pp 239–240

[85] Dumitrescu M, Toikkanen L, Sipilä P, Vilokkinen V, Melanen P, Saarinen M, Orsila S, Savolainen P, Toivonen M and Pessa M 2000 Modeling and optimization of resonant cavity light emitting diodes grown by solid source molecular beam epitaxy *Microelectron. Eng.* **51–52** 449–460

[86] Stevens R, Risberg A, Wurtemberg MV, Schatz R, Ghisoni M and Streubel K 1999 High-speed visible VCSEL for POF Data Links *POF Conference* (Japan, 1999)

[87] Streubel K and Stevens R 1998 250 Mbit/s plastic fibre transmission using 660 nm resonant cavity light emitting diodes *Electron. Lett.* **34** 1862–1863

[88] Wirth R, Karnutsch C, Kugler S and Streubel K 2001 High efficiency resonant-cavity leds emitting at 650 nm *Photon. Technol. Lett.* **13** 421

[89] De Neve H, Blondelle J, Daele P V, Demester P, Baets R and Borghs G 1997 Recycling of guided mode light emission in planar microcavity light emitting diodes *Appl. Phys. Lett.* **70** 799–801

[90] Yablonovitch E, Gmitter T J and Bhat R 1988 Inhibited and enhanced spontaneous emission from optically thin AlGaAs/GaAs double heterostructures *Phys. Rev. Lett.* **61** 2546–2549

[91] John S 1987 Strong localization of photons in certain disordered dielectric superlattices *Phys. Rev. Lett.* **58** 2486–2489

[92] Baba T 1997 Photonic crystals and microdisk cavities based on GaInAsP-InP system *J. Sel. Top. Quantum Electron.* **3** 808–830

[93] Yablonovitch E 1993 Photonic band-gap crystals *J. Physique* **5** 2443–2460

[94] Boroditsky M, Krauss T F, Coccioli R, Vrijen R, Bhat R and Yablonovitch E 1999 Light extraction from optically pumped light-emitting diode by thin-slab photonic crystals *Appl. Phys. Lett.* **75** 1036–1038

[95] Barnes W L, Kitson S C, Preist T W and Sambles J R 1996 Photonic surfaces *Microcavities Photonic Bandgaps: Physics Appl.* 265–274

[96] Evans G A *et al* 1991 Characterization of coherent two-dimensional grating surface emitting diode laser arrays during cw operation *J. Quantum Electron.* **27** 1594–1605

[97] Eriksson N, Hagberg M and Larsson A 1995 Highly efficient grating-coupled surface-emitters with single outcoupling elements *Photon. Technol. Lett.* **7** 1394–1396

[98] Erchak A A, Ripin D J, Fan S, Rakich P, Joannopoulos J D, Ippen E P, Petrich G S and Kolodziejski L A 2001 Enhanced coupling to vertical radiation using a two-dimensional photonic crystal in a semiconductor light-emitting diode *Appl. Phys. Lett.* **78** 563–565

[99] Rattier M 2001 Diodes électro-luminescentes à cristaux photoniques: extraction de la lumière guidée *PhD Thesis* (Ecole Polytechnique de Paris)

[100] Boroditsky M, Vrijen R, Krauss T F, Coccioli R, Bhat R and Yablonovitch E 1999 Enhancement from thin-film 2-D photonic crystals *J. Lightwave Technol.* **17** 2096–2112
[101] Barnes W L 1999 Electromagnetic crystals for surface plasmon polaritons and the extraction of light from emissive devices *J. Lightwave Technol.* **17** 2170–2182
[102] Wyszecki G and Stiles W S 1982 *Colour Science Concepts and Methods, Quantitative Data and Formula* 2nd edn (New York: Wiley)
[103] Schlotter P, Schmidt R and Schneider J 1997 Luminescence conversion of blue light emitting diodes *Appl. Phys. A* **64** 417–418
[104] Schlotter P, Baur J, Hielscher C, Kunzer M, Obloh H, Schmidt R and Schneider J 1999 Fabrication and characterization of GaN/InGaN/AlGaN double heterostructure LEDs and their application in luminescence conversion (LUCOLEDs) *Mater. Sci. Eng. B* **B59** 390–394
[105] Nakamura S, Senoh M, Iwasa N, Nagahama S, Yamada T and Mukai T 1995 Superbright green InGaN single-quantum-well-structure light-emitting diodes *Japan. J. Appl. Phys. Lett.* **34** L1332–L1335
[106] Nakamura S, Pearton S and Fasol G 1997 *The Blue Laser Diode* (Berlin: Springer)
[107] Baur J, Schlotter P and Schneider J 1998 White light emitting diodes *Adv. Solid State Phys.* **67** 67–78
[108] Bogner G, Debray A and Höhn K 2000 High performance epoxy casting resins for SMD-LED packaging *Light-Emitting Diodes: Research, Manufacturing, and Applications IV* (*Proc. SPIE vol 3938*) pp 249–261
[109] Crawford M H, Han J, Chow WW, Banas MA, Figiel JJ, Zhang LZ and Shul RJ 2000 Design and performance of nitride-based UV LEDs *Light-Emitting Diodes: Research, Manufacturing, and Applications IV* (*Proc. SPIE vol 3938*) pp 13–23
[110] Cree Lighting Inc. 2001 Press Release
[111] OSRAM, 1987 *Taschenbuch der Lampentechnik* (Berlin: Osram GmbH)
[112] Mueller-Mach R and-Mueller G O 2000 White light emitting diodes for illumination *Light-Emitting Diodes: Research, Manufacturing, and Applications* (*Proc. SPIE vol 3938*) pp 30–42
[113] *High-Brightness LED Market Review and Forecast* 2001 Strategies Unlimited
[114] *Commission Internationale de lEclairage Proceedings* 1931 (Cambridge: Cambridge University Press)
[115] Ryer A 1998 *Light Measurement Handbook* (International Light Inc.) http://www.intl-light.com

Appendix A

A.1 The CIE colour system

In 1931, the Commission Internationale de l'Eclairage (CIE) produced the 'Colour Standard Table' in order to create an objective method of determining colours [114]. As a basis for standardization, the CIE chose the response of the three sets of colour receptors in the eye. Colours are measured by comparison with an additive mixture of three elementary colours and specified by the tristimulus values X, Y and Z. The next task was to find a way for a two-dimensional representation of colours in a colour map, similar to geographical maps. This was achieved by calculating a new set of variables, the colour-masses x, y and z from the measured tristimulus values by dividing each of them by their total sum: $x = X/(X + Y + Z)$, $y = Y/(X + Y + Z)$ and $z = Z/(X + Y + Z)$. With this conversion, only two values, e.g. x and y, remain independent and can be used as coordinates in a two-dimensional chart. A representation with the x-values on the horizontal axis and the y-values on the vertical axis is called the CIE chromaticity diagram. The saturated colours are located on the locus of the chromaticity diagram. A straight line, the 'purple line' joins the red and violet ends of the spectral locus to form a closed diagram. The colours on the purple line are mixtures of the spectral colours 770 and 400 nm on the right- and left-hand side. All existing colours lie within the tongue shape delineated by the line of spectral colours and the purple line (figure A1).

The CIE defined the colour coordinates of several illuminants to describe commonly used white light sources (table A1).

The special point E with the coordinates $x = y = 1/3$ is denoted 'equal energy' and locates the achromatic point representing greys and white. Complementary colours lie on opposite sides of the achromatic point. An important definition for the colour of LEDs is the *dominant wavelength*. It is defined as the spectral colour that is perceived to be the same as the colour of the LED. In the chromaticity diagram, the dominant wavelength can be determined by drawing a straight line from the achromatic point through the point representing LED colour (figure A2). The intersection of the

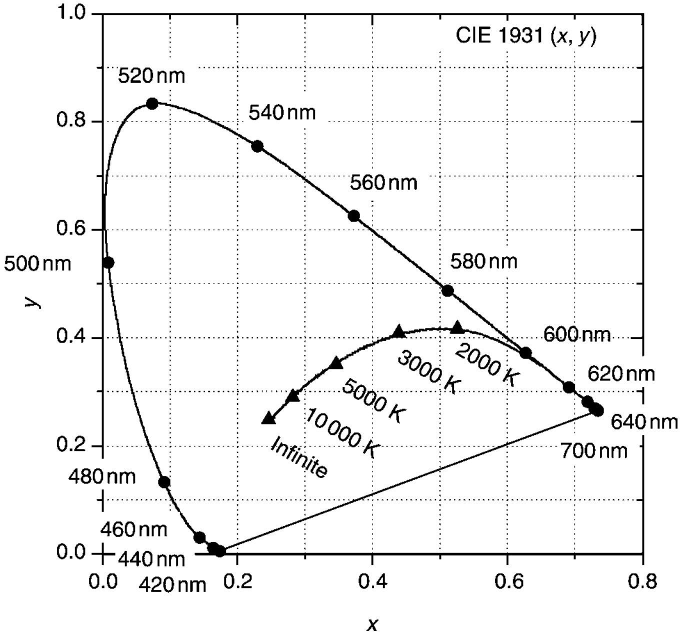


Figure A1. 1931 CIE diagram of colours. The solid line gives the black body radiation at temperatures from 1000 K to infinite.

prolonged line with the perimeter of the diagram gives the spectral colour that defines the dominant wavelength. This concept is also used to define the colour purity of a light source. *Colour purity* is a measure of how close the colour point lies to the perimeter of the CIE diagram. It is defined as the ratio of the distance of the colour point to the perimeter, weighted with the distance of the achromatic point to the same intersection point. Thus, the purity of a colour ranges from 0 for the achromatic point to 1 for saturated colours at the perimeter.

The human eye sensitivity has been defined and standardized by the CIE in 1924. Usually two response curves V_λ and V'_λ for day and night vision (photopic and scotopic vision) are used (figure A3). In daylight vision, yellowish-green light at 555 nm stimulates the eye more than blue or red light. At this wavelength, the human eye can detect about 10 photons s^{-1} or a radiant power of 3.58×10^{-18} W. At blue (450 nm) and red (650 nm), the eye detection limits are 214 and 126 photons s^{-1}, respectively [115].

Table A1. Colour coordinates and colour temperature of CIE Illuminants. Illuminant E is an imaginary colour with equal values for *x*, *y* and *z*.

	Illuminant	*x*	*y*	*T* (K)
Incandescent lamp	A	0.4476	0.4074	2856
Direct sunlight	B	0.3484	0.3516	4870
Overcast sunlight	C	0.3101	0.316	6770
Daylight	D_{65}	0.3128	0.32932	6504
Equal energy	E	0.3333	0.3333	

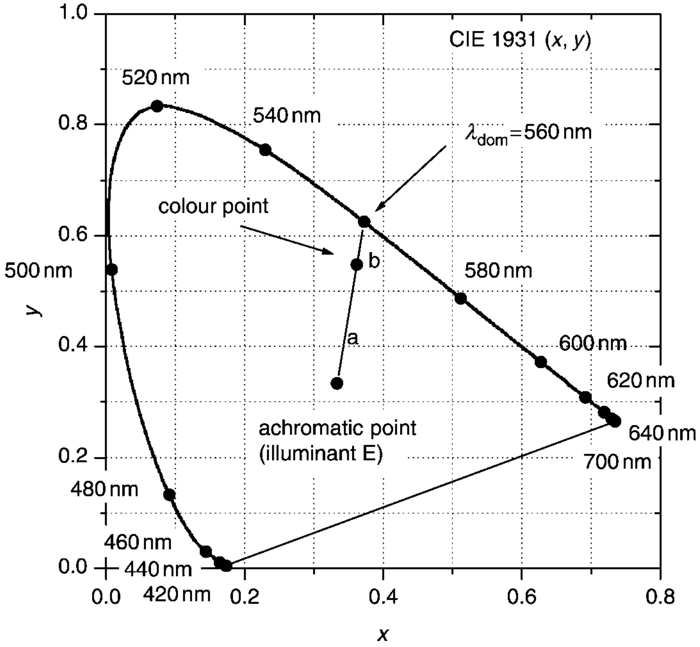


Figure A2. Dominant wavelength and spectral purity of a light source. The colour purity is defined as $a/(a+b)$.

A.2 Photometric and radiometric units

Radiometry is the measurement of electromagnetic radiation at all frequencies, whereas photometry is restricted to the measurement of visible light. The difference between the two methods is that in photometry everything is weighted by the human eye sensitivity.

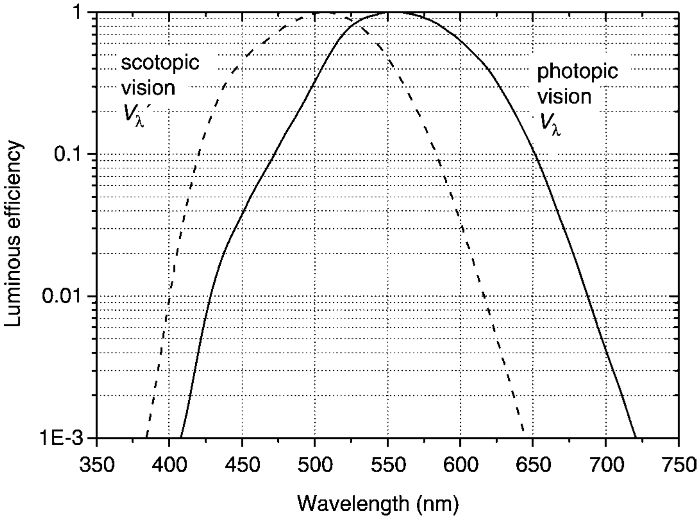


Figure A3. Spectral eye sensitivity curves V_λ and V'_λ for light-adapted (photopic) and dark-adapted (scotopic) vision. The peak of V'_λ is shifted 43 nm towards shorter wavelength.

Table A2. Radiometric and photometric units. Symbols for the radiometric quantities have the subscript e, and photometric quantities have the subscript v. The SI unit watt is suffixed with 'opt' and 'el' in order to denote the units of optical and electrical power.

Radiometric units		Photometric units	
Radiant flux ϕ_e	W_{opt} (watt)	Luminous flux ϕ_v	lm (lumen)
Radiant efficiency	W_{opt}/W_{el}	Luminous efficiency	lm/W_{el}
		Luminous efficacy	lm/W_{opt}
Radiant intensity I_e	W_{el}/sr	Luminous intensity I_v	lm/sr = cd (candela)
Radiance L_e	$W_{el}/sr\,m^2$	Luminance L_v	cd/m^2

W_{opt}, unit of the radiant flux (watt); W_{el}, unit of electrical power (watt); sr, SI unit of the solid angle (steradian).

In radiometry, the power flux of electromagnetic radiation is measured in watts (W). This flux is called the *radiant flux* ϕ_e of a light source. The photometric pendent to the radiant flux is the *luminous flux* ϕ_v and has the SI unit lumen (lm). The definition of lumen is related to the CIE eye response curve where 1 W of radiant flux at 555 nm wavelength is defined to have a luminous flux of 683 lm.

$$\phi_v = 683 V_\lambda \phi_e. \tag{A1}$$

Luminous efficiency is the ratio of luminous flux measured in lumen and the electrical power used to generate this flux measured in watt. One very confusing thing here is that watt as the general SI unit for power is used for optical and electrical measurements. Thus another term, the *luminous efficacy K* defined as the ratio of luminous flux and radiant flux has the same unit $\mathrm{lm\,W^{-1}}$.

$$K = \frac{\phi_v}{\phi_e}. \tag{A2}$$

Luminous efficiency is inevitably lower than the luminous efficacy, because some of the input power is lost in form of heat and does not appear as emitted radiant power. The efficiency of converting electrical power into radiant flux is measured by the *radiant efficiency* with the unit $\mathrm{W\,W^{-1}}$.

The total radiant or luminous flux does not indicate how the LED output varies with direction. For many applications, it is important to know how much power in watts or lumen is concentrated within a narrow angular range in a particular direction. This is measured by the *radiant intensity* I_e or the luminous intensity I_v of the LED in that direction. The units for *luminous intensity* are $\mathrm{W\,sr^{-1}}$ (sr = steradian) and $\mathrm{lm\,sr^{-1}}$, respectively. The latter unit, $\mathrm{lm\,sr^{-1}}$ is called candela (cd) and is equal to $1\,\mathrm{lm\,sr^{-1}}$.

The concept of radiant and luminous intensity applies only to point sources. Sources with an extended light-emitting area are characterized by the luminance, defined as luminous intensity per unit light-emitting area. The unit of *luminance* is $\mathrm{cd\,m^{-2}}$. Since light sources that emit the same luminous intensity from a smaller emitting area appear brighter, the term luminance is correlated with the qualitative term 'brightness'. The radiometric counterpart to luminance is *radiance*, which is the radiant flux per steradian per square metre of emitting surface area (table A2).

B1.2
Semiconductor lasers

Jens Buus

B1.2.1 Introduction

The semiconductor laser is now by far the most commonly used laser type, with several millions now being produced every month. It is also the most versatile laser type, allowing a variety of performance parameters to be optimized for specific applications: For example, the lasing wavelength can be tailored by the use of different materials and compositions, the spectral properties can be optimized, and lasers can be designed for high power levels.

Semiconductor lasers operating at visible wavelength around 650 nm (red) are widely used in pointers, bar code readers and DVD players. Lasers with wavelengths around 800 nm (just outside the visible part of the spectrum) are used in CD players and CD-ROM drives. The lasers for these applications have comparatively simple structures, making it possible to manufacture packaged laser assemblies for prices of the order of $1. Lasers used as transmitters in fibre optic communication systems, on the other hand, can cost several thousand dollars. These lasers usually work at wavelengths around 1300 nm (where the dispersion in standard fibres has its minimum) or 1550 nm (where the fibre loss has its minimum). Devices with a variety of performance characteristics, and prices, are available for different types of communication systems.

The widespread use of semiconductor lasers is due to a number of inherent advantages, including:

- simple electrical pumping merely by passing a current through the device;
- possibility of direct modulation by varying the pump current;
- high efficiency, over 50% conversion efficiency has been achieved;
- robustness and reliability;
- small size, typically 0.3 mm by 0.3 mm, i.e. a density of about 1000 per cm^2 is possible;
- amenable to automated mass production;
- large degree of freedom for optimization for specific applications.

The semiconductor laser differs from other laser types in many respects. A main reason for these differences is that the lasing transition occurs between electronic states in relatively broad *energy bands*, rather than between *discrete energy levels*. Another important difference is that very high gain coefficients (well over $100\,\mathrm{cm}^{-1}$) are possible. This in turn means that the laser cavities can be very short, and the end reflectivities can be quite low.

In this chapter, the basic principles and properties of semiconductor lasers will be outlined (sections B1.2.2–B1.2.6), and examples of some of the various semiconductor laser types will be given (sections B1.2.7–B1.2.12). The various sections are intended to be relatively independent, and the sections on the specific examples can be read without studying the more theoretical parts (sections B1.2.4–B1.2.6 in particular) in detail. For papers describing the historical development of semiconductor lasers, the reader is referred to [1, 2].

B1.2.2 General principles and basic structures

Nearly all semiconductor lasers are based on the *double heterostructure*. (We note in passing that Alferov and Kroemer shared half the 2000 Nobel Prize in physics for their work on heterostructures.) In this structure, a material with a relatively narrow bandgap—*the active layer*—(normally undoped) is sandwiched between a pair of n-type and p-type materials with wider bandgaps—*the confinement layers*. When this structure is forward biased, *quasi-Fermi levels* are formed, and electrons and holes are injected into the active layer from the n-type and p-type material, respectively. The Fermi level(s) is the energy where the probability of electron or hole occupation is 50%, and this (these) level(s) determine the energy distribution of electrons in the conduction band and holes in the valence band. *Population inversion*, and thereby gain, is achieved when the quasi-Fermi level separation exceeds the bandgap of the active layer, see figure B1.2.1.

If the bandgap difference is sufficiently large $E_{g_2} - E_{g_1} \gg k_B T$, where k_B is Boltzmann's constant and T the temperature. Then carriers injected into the active layer cannot escape over the heterobarrier, and carrier recombination can only take place in the active layer. Light generated in the active layer is not absorbed in the confinement layers, because semiconductors are transparent to light with a photon energy lower than the bandgap. The photon energy E is given by the product of Planck's constant h and the optical frequency ν, which is in turn equal to the speed of light in vacuum, c, divided by the wavelength, λ. Hence

$$E = h\nu = \frac{hc}{\lambda}. \tag{B1.2.1}$$

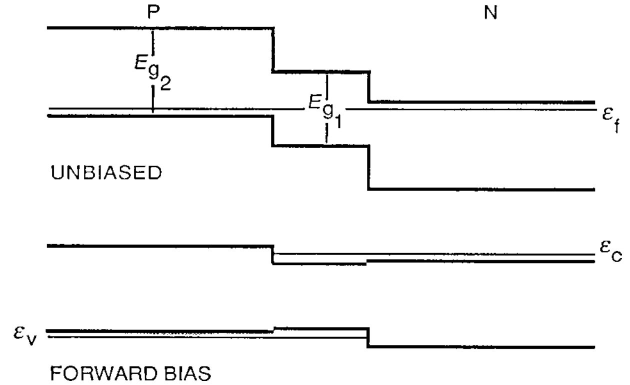


Figure B1.2.1. Double heterostructure formed in a material with a narrow bandgap E_{g_1}, placed between n-type and p-type materials with a wider bandgap E_{g_2}. Carriers cannot have energies that correspond to states within the bandgap. Without bias, the Fermi level ε_f is continuous. Under forward bias quasi-Fermi levels ε_c and ε_v are formed for electrons in the conduction band and for holes in the valence bands, respectively.

At a given optical frequency, a narrow bandgap semiconductor generally has a higher refractive index than a semiconductor with a wider bandgap. Consequently, the structure shown in figure B1.2.1 also forms a planar dielectric waveguide, with a high index core between a pair of low index cladding layers, analogous to an optical fibre. It is characteristic for a dielectric waveguide that only a part of the optical power is confined to the core, because the power distribution extends well into the cladding layers. The power fraction in the core is known as the *confinement factor*, and denoted by the symbol Γ.

The optical field distribution supported by the waveguide is known as a *mode*. The optical field propagates at a speed given by c/n_{eff}, where the *effective index*, n_{eff}, of the mode is higher than that of the cladding layers, but lower than that of the core.

The laser cavity forms a resonator, and the cavity length L and the effective index n_{eff} are related to the lasing wavelength λ by

$$n_{\mathrm{eff}}L = M\frac{\lambda}{2}. \tag{B1.2.2}$$

This states that the optical length of the cavity is an integer number of half wavelengths, where $M \gg 1$ is known as the (longitudinal) *modenumber*. The separation, also known as the *modespacing*, between two wavelengths (*longitudinal modes*) satisfying this condition (corresponding to modenumbers M and $M+1$) is

$$\Delta\lambda = \frac{\lambda^2}{2n_{\mathrm{eff}}L}. \tag{B1.2.3}$$

With an effective index of 3.5 and a cavity length 300 μm, the modespacing is about 0.2 nm for a lasing wavelength of 650 nm, and about 1 nm for a lasing wavelength of 1550 nm. Strictly, it is the *effective group index* which should be used in equation (B1.2.3) in order to account for dispersion.

The combination of amplification and feedback from the mirrors forms an oscillator, and oscillation takes place if the amplification balances the loss caused by light escaping through the mirrors. With a gain factor g, a cavity length L, and two mirrors both with a power reflectivity of R, this condition can be written as

$$\exp(gL)R\,\exp(gL)R = 1. \tag{B1.2.4}$$

Gain levels can be very high, this means that the lasing condition, as expressed in equation (B1.2.4), can be satisfied for low values of the reflectivity R, even for short cavity lengths (e.g. 300 μm as mentioned earlier). As semiconductors typically have refractive index values around 3.5, sufficient reflectivity (about 30–35%) can be obtained from a cleaved facet, and there is no need for special high reflectivity external mirrors as is the case for most other laser types.

It is important to note that the gain factor g in equation (B1.2.4) is the *net modal gain* experienced by the lasing mode. Taking internal optical losses, given by the *internal loss factor* α_{int}, into account, and noting that only the power fraction in the active layer, given by the confinement factor Γ, experiences gain, we have

$$g = \Gamma g_{\mathrm{act}} - \alpha_{\mathrm{int}} \tag{B1.2.5}$$

where g_{act} is the gain in the active layer.

If the two facet reflectivities are equal, as in equation (B1.2.4), equal amounts of power will be emitted from the two ends. This power balance can be changed by applying a facet coating to either increase the reflectivity (HR—high reflection coating), or to decrease the reflectivity (AR—anti-reflection coating). If, for example, the back facet reflectivity is increased to near unity, nearly all the light will be emitted from the front facet.

Light emitted from the ends can be accounted for by introducing an *end loss factor* α_{end}, which can be found by solving equation (B1.2.4) for the gain required to overcome the end losses. For the case of two equal facet reflectivities, we get

$$\alpha_{\text{end}} = \frac{1}{L}\ln\left(\frac{1}{R}\right). \tag{B1.2.6}$$

The condition for lasing is that the *available* gain, given by equation (B1.2.5), equals the gain *required* for lasing, given by equation (B1.2.6). This *lasing condition* can be written as

$$g = \Gamma g_{\text{act}} - \alpha_{\text{int}} = \alpha_{\text{end}}. \tag{B1.2.7}$$

The gain in the active layer, g_{act}, is an increasing function of the carrier density N in the active layer. As long as the carrier density is so low that the gain is insufficient to satisfy the lasing condition, the carrier density is related to the current I supplied to the laser by

$$I = \frac{eVN}{\tau} \tag{B1.2.8}$$

where e is the unit charge, V the volume of the active region and τ the *carrier lifetime*. (We should note here that the carrier lifetime may describe several recombination mechanisms, some of which in turn depend on the carrier density. Hence the carrier lifetime cannot necessarily be considered as a simple constant.)

We can now see that the laser is a threshold device. For low values of the current I, there is insufficient gain to satisfy the lasing condition, and only a small amount of light is generated by spontaneous emission. Lasing starts at the *threshold current* I_{th}, which is the current that gives a carrier density N_{th} that satisfies the lasing condition

$$\Gamma g_{\text{act}}(N_{\text{th}}) = \alpha_{\text{end}} + \alpha_{\text{int}} = g_{\text{th}} \tag{B1.2.9}$$

where N_{th} is referred to as the *threshold carrier density* and g_{th} is the *threshold gain*. For currents above the threshold, the gain does not increase any further and the carrier density remains at the threshold value. Stimulated recombination now takes place, and the amount of optical power emitted from the laser increases sharply. We describe this by introducing the *differential efficiency* η_{d}, which is the ratio of light 'lost' by emission through the facets to the total loss, hence

$$\eta_{\text{d}} = \frac{\alpha_{\text{end}}}{\alpha_{\text{end}} + \alpha_{\text{int}}}. \tag{B1.2.10}$$

The dependence of the carrier density and the emitted laser power as functions of the laser current are shown in figure B1.2.2.

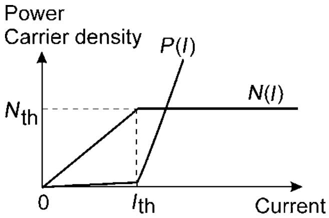


Figure B1.2.2. Carrier density N, and emitted laser power P, as functions of the current supplied to the laser. Below the threshold current there is only spontaneous emission, above the threshold the carrier density (and hence gain) remains fixed, and the power increases.

In order to work out the total power balance, we consider a laser operated at a current I, with an electrical resistance R', and a photon energy $h\nu$. Introducing the unit charge e, the voltage drop U across the laser is given by

$$U = IR' + \frac{h\nu}{e}. \tag{B1.2.11}$$

By multiplying this equation by the current I, we can write the power supplied to the laser as

$$IU = I^2R' + I_{\mathrm{th}}\frac{h\nu}{e} + (I - I_{\mathrm{th}})(1 - \eta_{\mathrm{d}})\frac{h\nu}{e} + (I - I_{\mathrm{th}})\eta_{\mathrm{d}}\frac{h\nu}{e}. \tag{B1.2.12}$$

The four terms on the right-hand side represent, respectively: ohmic losses, recombination losses, internal optical losses and emitted laser power (with one half of this power emitted from each end if the reflectivities are equal).

Example: For a facet reflectivity of 0.35 and a cavity length of 300 μm, the end loss, from equation (B1.2.6), is 35 cm^{-1}. If the internal optical loss is 20 cm^{-1}, and the confinement factor is 0.2, then equation (B1.2.9) gives the gain required in the active region as 275 cm^{-1}, and according to equation (B1.2.10), the differential efficiency is 0.64 (i.e. 0.32 per facet). We next assume that the width of the active region is 2 μm and its thickness is 0.1 μm. This gives an active volume of $60 \times 10^{-18}\ \mathrm{m}^3$. If the threshold carried density is $2 \times 10^{24}\ \mathrm{m}^{-3}$ and if the carrier lifetime is 2 ns, then, according to equation (B1.2.8), the threshold current is 9.6 mA.

As we have seen, the double heterostructure confines both carriers and light in the vertical direction. In order to reduce the volume of the active region (and hence, according to equation (B1.2.8), reduce the threshold current), there is also a need for lateral carrier confinement. The simplest approach is to limit the current injection to a narrow stripe along the length of the laser. However, there are problems associated with this: Firstly, there is still no well-defined carrier confinement, and, due to current spreading in the confinement layer and carrier diffusion in the active layer, the carriers will occupy a region much wider than the stripe, effectively giving a larger active volume. Secondly, there is no well-defined lateral optical waveguide, resulting in reduced beam quality and erratic optical behaviour.

These problems are partly solved by using the *ridge waveguide* (RW) structure shown in figure B1.2.3. The current is injected into a narrow ridge, and, due to the reduced thickness of the confinement layer outside the ridge, current spreading is reduced. In addition, the presence of the ridge creates a weak lateral optical waveguide, thus improving the optical confinement properties.

Strong lateral confinement of both carriers and light requires the active region to be completely surrounded by material with a larger bandgap. This is achieved in the *buried heterostructure* (BH) shown in figure B1.2.4. The improved performance of the BH lasers comes at a cost of a more complicated

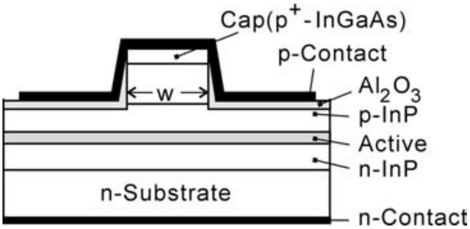


Figure B1.2.3. Schematic cross-section of an InP-based ridge waveguide laser. Due to current spreading, the effective active volume is somewhat wider than that corresponding to the ridge width w.

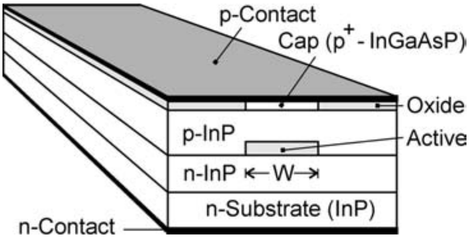


Figure B1.2.4. Schematic diagram of an InP-based buried heterostructure laser.

fabrication process. In order to fabricate a BH structure, multiple epitaxial growths, with intermediate stripe definition and etching processes, are necessary.

B1.2.3 Materials and fabrication technologies

Until now, almost all semiconductor lasers have been based on materials with a direct bandgap, usually using GaAs or InP substrates. In order to fabricate a double heterostructure, it is necessary to find materials that can be grown on these substrates, which have bandgaps different from the substrate material, and which have a lattice constant compatible with the substrate.

The simplest example is $Al_xGa_{1-x}As$, which consists of two group III elements (Al and Ga) and one group V element. The bandgap increases with the Al fraction (x), but the lattice constant remains nearly unchanged. A constant lattice constant is important in order to avoid the formation of defects during the growth of the material. A double heterostructure semiconductor laser is formed by having an active layer with a low Al content, and confinement layers with a high Al content. The difference in Al content must be sufficiently high to ensure sufficient carrier confinement. The photon energy for the laser is slightly higher than the bandgap of the active layer, and by changing the Al fraction in the active layer, this photon energy can be changed, resulting in lasers with wavelengths in the 800–900 nm range. We note that because the photon energy is higher than the bandgap of the substrate (GaAs), the lower cladding layer must be sufficiently thick to avoid absorption losses in the substrate.

More design flexibility is possible by using two group III and two group V materials. The prime example is $In_{1-x}Ga_xAs_yP_{1-y}$. Lattice matching to an InP substrate is achieved by having $x \approx 0.47y$, and the second degree of freedom in the composition can be used to vary the bandgap, and hence the wavelength. This makes it possible to fabricate lasers with wavelengths from around 1100 nm to nearly 1700 nm, covering the important 'telecommunications' wavelengths around 1300 and 1550 nm. Figure B1.2.5 shows the relation between bandgap and lattice constant. An alternative '*long wavelength*' material, also lattice matched to InP, is $(Al_xGa_{1-x})_yIn_{1-y}As$.

Lasing at visible wavelengths is convenient for applications such as bar code readers and pointers, and access to shorter wavelengths, with a smaller focused spot size potential, also makes it possible to achieve a higher information density on storage media such as DVDs. Lasers operating in the 600–700 nm range (red) can be fabricated using $(Al_xGa_{1-x})_yIn_{1-y}P$ compounds on GaAs substrates.

For even shorter wavelengths, corresponding to blue and green, ZnSe-based materials have been used, and there is currently a high level of activity on the development of short wavelength lasers based on $(Al_xGa_{1-x})_yIn_{1-y}N$ compounds.

For wavelengths around 2 μm antimony (Sb) containing materials have been used, and for even longer wavelengths there has been work on 'lead-salts' such as Pb(SSe), (PbSn)Te and (PbSn)Se.

It should be noted that, by using one or more very thin active layers, the lattice matching requirement can be relaxed making it possible to get access to an extended range of wavelengths.

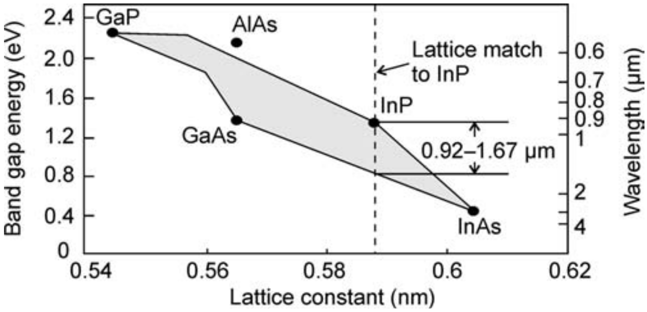


Figure B1.2.5. Bandgap energy, corresponding wavelength and lattice constant for InGaAsP compounds (and AlAs). Lattice matching to InP is indicated by the broken line. Notice the similarity of the lattice constants for GaAs and AlAs.

One example is lasers fabricated on GaAs substrates, with (InGa)As active layers and (AlGa)As confinement layers, operating at a wavelength of 980 nm for use as pump sources for Er-doped fibre amplifiers. This topic is discussed further in sections B1.2.5 and B1.2.9.

The growth and processing of layers with sub micrometre thickness, with high precision and a high degree of uniformity is technologically very demanding. The most commonly used technique for growth of heterostructures is *metal-organic chemical vapour phase epitaxy* (MOVPE). Hydrides such as arsine (AsH_3), phosphine (PH_3) and organometallics such as tri-methyl-gallium $Ga(CH_3)_3$ and tri-methyl-indium $In(CH_3)_3$ are carried by hydrogen and react on the surface of the wafer. The material composition is controlled by adjusting the flow rate of the various sources. Large wafers can be grown, and some reactors allow multi-wafer handling, making this technique suitable for large volume manufacturing. MOVPE requires very stringent safety measures due to the toxicity of the hydrides.

An alternative technology is *molecular beam epitaxy* (MBE). In this technique, beams of atoms or molecules impinge on a heated substrate. MBE is particularly suitable for high precision growth of very thin layers, but the requirement for ultra-high vacuum makes it very expensive. The fabrication of semiconductor lasers also requires high precision lithography and etching to form the active stripe, and some laser structures require multiple growths.

B1.2.4 Basic theory for gain in semiconductor lasers

The detailed calculation of the gain in the active layer of a double heterostructure is a rather lengthy procedure, and here we will only consider the main points of that procedure. For a more comprehensive and more accurate treatment the reader is referred to one of the standard texts, for example, reference [2] on the further reading list. Readers not familiar with basic solid state physics concepts may skip this section as it is not essential for the understanding of the remainder of this chapter.

In a semiconducting material, the electrons and holes do not move as free particles, so the relation between energy, momentum and mass is different from that of a free particle. However, in many cases, the relation between energy and momentum is still nearly parabolic, hence an *effective mass* can be defined. In most of the materials of interest for semiconductor lasers, we find that the effective mass for the electrons in the conduction band, m_c, is smaller than the effective mass for the holes in the valence band, m_v.

We consider the recombination of a hole and an electron, as shown in the band diagram in figure B1.2.6. Note that conservation of momentum corresponds to a vertical transition in such a diagram.

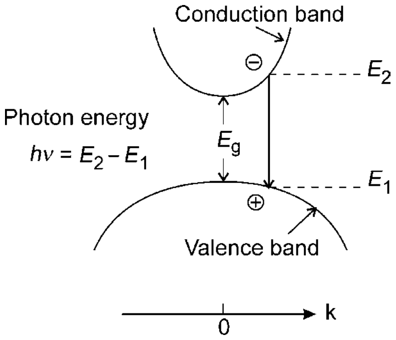


Figure B1.2.6. Band diagram of a semiconductor, where the effective mass for the electrons is smaller than the effective mass for the holes.

The energy of the photon emitted by the transition shown in figure B1.2.6 is given by

$$h\nu = E_2 - E_1 = E_\mathrm{g} + \frac{\hbar^2 k^2}{2m_\mathrm{c}} + \frac{\hbar^2 k^2}{2m_\mathrm{v}} = E_\mathrm{g} + \frac{\hbar^2 k^2}{2m_\mathrm{r}} \tag{B1.2.13}$$

where k is the wave vector, $\hbar k$ the momentum and $\hbar$ Planck's constant divided by 2π and the last equality serves as the definition of the *reduced mass* m_r. For vertical (i.e. momentum conserving) transitions we have, with $E = 0$ at the bottom of the conduction band

$$E_2 = (h\nu - E_\mathrm{g})\frac{m_\mathrm{r}}{m_\mathrm{c}} \quad \text{and} \quad E_1 = -E_g - (h\nu - E_\mathrm{g})\frac{m_\mathrm{r}}{m_\mathrm{v}}. \tag{B1.2.14}$$

We next need to know how many states are available corresponding to a given energy. This is given by the *density of states function,* which, for the case of parabolic bands, has a simple dependence on effective mass and energy. For the conduction band, we have

$$D_\mathrm{c}(E_2) = \frac{\sqrt{2}}{\pi^2 \hbar^3} m_\mathrm{c}^{3/2} \sqrt{(h\nu - E_\mathrm{g})\frac{m_\mathrm{r}}{m_\mathrm{c}}} \tag{B1.2.15a}$$

and for the valence band

$$D_\mathrm{v}(E_1) = \frac{\sqrt{2}}{\pi^2 \hbar^3} m_\mathrm{v}^{3/2} \sqrt{(h\nu - E_\mathrm{g})\frac{m_\mathrm{r}}{m_\mathrm{v}}}. \tag{B1.2.15b}$$

The *reduced density of states* at the transition energy is the reciprocal of the sum of the reciprocal densities of states given in equations (B1.2.15a) and (B1.2.15b)

$$D_\mathrm{r}(h\nu) = \frac{\sqrt{2}}{\pi^2 \hbar^3} m_\mathrm{r}^{3/2} \sqrt{(h\nu - E_\mathrm{g})}. \tag{B1.2.16}$$

Electrons and holes are *fermions*, and the probability that a certain energy state is occupied is therefore given by a *Fermi–Dirac distribution*. The number of occupied states with a given energy in the conduction band (unoccupied states, i.e. holes, in the valence band) is given by the product of the density of states in the conduction (valence) band, multiplied by the probability of occupation f_c (the probability of finding an empty state is $1 - f_v$). These probabilities are given by

$$f_c(E) = \frac{1}{1 + \exp\left(\frac{E - \varepsilon_c}{k_B T}\right)} \quad \text{and} \quad 1 - f_v(E) = \frac{1}{1 + \exp\left(-\frac{E - \varepsilon_v}{k_B T}\right)} \tag{B1.2.17}$$

where the Fermi level ε_c (ε_v) (cf figure B1.2.1) is found by requiring that the integral (over all energies) of the product of the density of states and the occupation probability is equal to the total number of electrons (holes). In the case of charge neutrality and no doping, the number of electrons must be equal to the number of holes. In equation (B1.2.17) k_B is Boltzmann's constant.

We can now write an expression for the increase of the optical energy density ρ, per unit time, at a given photon energy due to stimulated emission. This expression is the product of the reduced density of states, the occupation probability for the conduction band, the probability of having an empty state in the valence band and the transition probability (the Einstein B coefficient)

$$\frac{d\rho}{dt}(h\nu)_{\text{stim}} = D_r(h\nu) f_c(E_2)(1 - f_v(E_1))B\rho. \tag{B1.2.18}$$

The expression for absorption is similar, but the Fermi–Dirac factors have opposite roles, and there is a negative sign because absorption decreases the optical energy density

$$\frac{d\rho}{dt}(h\nu)_{\text{abs}} = -D_r(h\nu) f_v(E_1)(1 - f_c(E_2))B\rho. \tag{B1.2.19}$$

By adding these two expressions, we get the result for the net increase in the optical energy density per unit time

$$\frac{d\rho}{dt}(h\nu)_{\text{net}} = D_r(h\nu)(f_c(E_2) - f_v(E_1))B\rho. \tag{B1.2.20}$$

The factor preceding ρ on the right-hand side of equation (B1.2.20) is simply the gain per unit time.

It is simple to show that, in order to have a positive gain factor, the separation between the Fermi levels must exceed the transition energy, which in turn must exceed the bandgap

$$f_c(E_2) > f_v(E_1) \text{ for } \varepsilon_c - \varepsilon_v > h\nu = E_2 - E_1 > E_g. \tag{B1.2.21}$$

The transition probability B, can be found from *Fermi's Golden Rule*, which gives the probability in terms of the matrix element of the interaction Hamiltonian

$$B = \frac{\pi}{2\hbar}|H|^2. \tag{B1.2.22}$$

It now requires some more detailed solid state physics theory to find the matrix element, here the result is simply quoted, m_0 is the free electron mass, ε_0 the free space permittivity and n the refractive index

$$|H|^2 = \frac{2e^2\hbar^2}{m_0^2 \varepsilon_0 n^2 h\nu}|M|^2 \tag{B1.2.23}$$

with the approximate expression

$$|M|^2 \approx \frac{m_0^2}{12m_c} E_g. \tag{B1.2.24}$$

Finally, the gain per unit length is written as the gain per unit time divided by the group velocity of the light in the laser, given by c divided by n_g. The resulting expression consists of three parts. The first is material dependent but energy independent, the second contains the square root of the energy dependence from the density of states, and the last factor comes from the occupation probabilities

$$g \approx \frac{\sqrt{2}e^2 m_r^{3/2} n_g}{12\pi\hbar^2 m_c \varepsilon_0 n^2 c} \frac{E_g}{h\nu} \sqrt{h\nu - E_g}(f_c - f_v). \tag{B1.2.25}$$

It must be stressed that this gain calculation is somewhat simplified, for example, we have only considered vertical transitions in the energy diagram, i.e. strict conservation of momentum, or k. Nevertheless, the result is of the correct magnitude, and gives a nearly correct parameter dependence.

The gain depends on the carrier (i.e. electron and hole) density through the Fermi–Dirac factor $(f_c - f_v)$. With increasing carrier density, the Fermi levels ε_c and ε_v move further into the bands, and the function $(f_c - f_v)$ moves towards higher energies, as illustrated in figure B1.2.7.

The gain is positive for photon energies in the interval between E_g and $(\varepsilon_c - \varepsilon_v)$. For higher photon energies, the gain is negative, i.e. the material is absorbing, and for photon energies below E_g the material is transparent. We notice that the photon energy corresponding to the maximum gain increases with increasing carrier density (this is known as *bandfilling*).

It is a characteristic of semiconductor lasers that the gain can reach values of several hundred cm^{-1}, far higher than in any other laser type. This explains why the lasers can be short, yet still lase with low facet reflectivities, cf equation (B1.2.6) for the facet loss. Another characteristic feature of the gain curves is that they are very wide compared to other laser types. This is due to the fact that the lasing transition occurs between two bands of energies rather than between two discrete energy levels.

The expression for gain given by equation (B1.2.25) is too cumbersome for direct use in device design and modelling. It is often sufficient to describe the gain, at a given photon energy, as a simple linear function of the carrier density

$$g \approx a(N - N_0). \tag{B1.2.26}$$

The parameter a is called the *gain slope* and N_0 is referred to as the *transparency carrier density*.

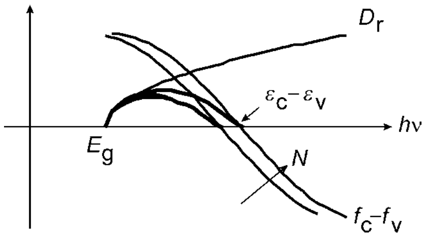


Figure B1.2.7. Fermi–Dirac factor $(f_c - f_v)$ for two different carrier densities. According to equation (B1.2.25), the gain is (apart from a numerical factor) mainly given by the product of $(f_c - f_v)$ and the reduced density of states D_r, which in turn varies as the square root of $(h\nu - E_g)$. This gives the gain curves for the two carrier densities, shown as bold curves.

In some cases, in particular, for quantum well lasers (see section B1.2.5), it is more appropriate to approximate the dependence of the gain on the carrier density by a logarithmic expression

$$g \approx g_0 \ln\left(\frac{N}{N_0}\right). \tag{B1.2.27}$$

B1.2.5 Structures with reduced dimensionality

In a semiconductor heterostructure, which is small in one or more dimensions, the carriers (electrons and holes) do not behave like particles. Instead, they start to display wave nature, and their behaviour must be treated according to the rules of quantum mechanics. By 'small' we mean that the size of the structure, in one or more directions, is of the order of the *de Broglie wavelength*, λ_d, which is given by

$$\lambda_\mathrm{d} = \frac{2\pi}{k} = \frac{h}{p} \tag{B1.2.28}$$

where k is the wave vector, h Planck's constant and p the momentum of the particle.

The case where the structure is 'small' in only one dimension, usually the thickness of the active layer, is known as a *quantum well*, and is one of the basic examples of the application of quantum mechanics. For a particle in an *infinitely deep* quantum well, the values for the wave vector are quantized according to

$$k_N = \left(\frac{\pi}{L_z}\right)N, \quad N = 1, 2, 3, \ldots \tag{B1.2.29}$$

where L_z is the *width* of the quantum well. By combining equations (B1.2.28) and (B1.2.29) with the relation between energy and momentum, we find that, with m being the mass (i.e. effective mass in the case of a semiconductor), the corresponding energy levels are given by

$$E_N = \frac{1}{8m}\left(\frac{hN}{L_z}\right)^2. \tag{B1.2.30}$$

For the case of a finite well depth, there are a finite number of energy levels, also known *as mini bands or sub bands*. If the well is sufficiently narrow (about a few nanometres), there will only be one possible energy level. For a more detailed treatment of quantum wells, the reader should consult, for example, reference [4] on the further reading list.

We note in passing that a quantum well is mathematically analogous to a waveguide: The different energy levels correspond to the modes of the waveguide; a 'small' waveguide supports one mode only; an infinitely deep quantum well is equivalent to a metallic waveguide.

Although a quantum well confines the carriers in one dimension, i.e. the carriers cannot move freely in the direction perpendicular to the well, there is no confinement in the other two dimensions, i.e. in the well plane, and the carriers can move freely in these two dimensions. The confinement changes the density of states function. We recall from equations (B1.2.15a) and (B1.2.15b) that, for bulk material the density of states varies with the square root of the energy

$$D(E) \propto \sqrt{E}. \tag{B1.2.31}$$

In a quantum well, the density of states associated with energy level N is given by a step function

$$D_N(E) = 0 \quad \text{for } E < E_N, \quad D_N(E) = \frac{m}{\pi\hbar^2 L_z} \quad \text{for } E > E_N \tag{B1.2.32}$$

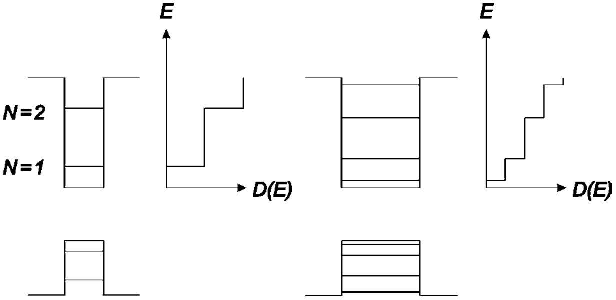


Figure B1.2.8. Energy levels in a 'thin' (left) and in a 'thick' (right) quantum well. Both the conduction bands and the valence bands are shown, assuming that the confinement energy is higher for the conduction bands than for the valence bands. It is also assumed that the effective mass is higher for the valence band than for the conduction band, leading to smaller energy level separations in the valence band. Because the quantum well has a finite depth, the positions of the energy levels deviate somewhat from the simple expression in equation (B1.2.30). The density of states functions for the conduction bands are also shown. The diagrams are schematic, and do not refer to a specific set of material parameters.

where again L_z is the width of the quantum well. As the value of the gain depends on the density of states function, cf equation (B1.2.20), the gain properties of a quantum well differ from those of bulk material. Very high gain coefficients are possible, but the gain tends to depend sub-linearly on the carrier density, as for example described by the approximation in equation (B1.2.27). In working out the details of the gain, a number of factors must be considered: All the subbands must be included, selection rules apply, so that transitions from a given subband in the conduction band are only allowed to the corresponding subband in the valence band, and the gain depends on the polarization of the optical field, see references [2, 4] on the further reading list for details.

According to equation (B1.2.30), the positions of the energy levels vary as the square of the level number. For a 'very thick' quantum well, the density of states summed over all the sub bands will therefore vary as the square root of the energy, thus recovering the result for bulk material given by equation (B1.2.31).

Figure B1.2.8 shows, schematically, examples of energy levels and density of states in quantum wells.

A single quantum well active region is too thin to provide efficient optical confinement, and it is normal to use a *separate confinement* structure, where the optical guiding is provided by a region which is much thicker than the quantum well. It is also possible to use more than one quantum well (see figure B1.2.9).

In a separate confinement structure, some of the important design considerations are de-coupled. For a given width of the quantum well(s), the optical confinement factor, defined as the overlap of the quantum well(s) and the optical field, is nearly completely determined by the design of the optical confinement region and almost independent of the quantum well(s). The quantum well properties, on the other hand, are determined by the combination of well width and depth. As the well width and depth determine the energy levels, these parameters partly determine the lasing wavelength.

A quantum well is very thin, typically less than 10 nm, and, although the gain in a quantum well can be high, the confinement factor is low, usually not more than a few percent per well. According to equation (B1.2.5) this gives a low value for the *modal* gain. As the gain increases sub-linearly with carrier density, it may be advantageous to have several wells, in spite of the fact that this increases the total

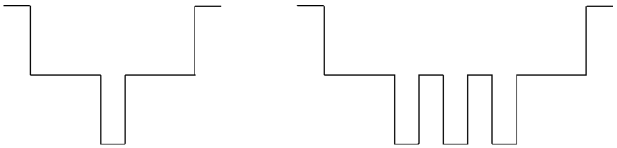


Figure B1.2.9. Band diagram (conduction band only) for separate confinement single (left) and multi (right) quantum well laser structures. Optical confinement is provided by the wide, low bandgap (high refractive index), optical confinement region. Carrier confinement is provided by the quantum well(s).

active volume. It is thus seen that, for a given threshold gain, it is possible to minimize the threshold current by optimizing the number of wells.

The degree of design freedom for the gain function in a quantum well structure also gives some scope for optimizing lasers for high-speed operation (see section B1.2.6 on laser dynamics).

We have previously mentioned that lattice matching is important to ensure a sufficient material quality. For a sufficiently thin layer, this condition is relaxed, because the misfit can be accommodated by strain without the formation of dislocations. For a single quantum well, the *critical thickness*, which is the maximum layer thickness that can be formed without dislocations, is approximately given by

$$t_\mathrm{c} \approx 0.2\left(\frac{\Delta a}{a}\right)^{-1}\ \mathrm{nm} \tag{B1.2.33}$$

where a is the lattice constant and Δa the lattice constant difference. As an example, this implies that a layer thickness of 20 nm is possible with a lattice mismatch of 1%. The lattice mismatch is accommodated by strain (tensile or compressive) of the quantum well layer.

The use of *strained quantum wells* makes it possible to use materials for which binary substrates are not available. An important practical result of this is that it gives access to new lasing wavelengths such as 980 nm, which is used for the pumping of fibre amplifiers (see also section B1.2.9). The strain also changes the band structure and hence the electronic properties of the material, and gives more flexibility for the optimization of laser performance.

By reducing the feature size(s) in one or two additional dimensions, *quantum wires* and *quantum dots* are formed. An important application of quantum dots is the fabrication of lasers operating at the 'telecom' wavelength of 1300 nm on GaAs substrates by using highly strained (InGa)As dots. In a quantum dot, the energy levels are discrete and the density of states function becomes a sum of δ functions. Consequently, some properties of these lasers become akin to those of gas or solid state lasers.

B1.2.6 Dynamics and noise

As mentioned in section B1.2.1, one of the major advantages of semiconductor lasers is that they can be directly current pumped. This also opens the possibility for *direct modulation* simply by varying the pump current. *Analogue modulation* can be performed by DC biasing the laser above threshold and superimposing an AC modulation current. In *digital modulation,* the laser is DC biased close to threshold, and a digital modulation current is superimposed.

In order to assess the modulation properties, we need to look at the dynamic behaviour of the laser, this behaviour is determined by the interaction of the carriers and the photons in the laser. As carriers (electrons and holes) are injected into the active region, they can either recombine spontaneously or by stimulated recombination brought about by the photon density in the active region. The photon density, on the other hand, is subject to both gain, due to the stimulated recombination of the carriers, and to losses, either internal losses, or losses due to photons being emitted from the end facets of the laser.

The interactions between the carrier density, N, and the photon density, S, are described by the so-called *rate equations* for the laser. In their simplest form, these equations can be written as

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{I}{eV} - \frac{N}{\tau} - v_\mathrm{g}\Gamma g_\mathrm{act}S \tag{B1.2.34}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = v_\mathrm{g}\Gamma g_\mathrm{act}S - v_\mathrm{g}(\alpha_\mathrm{end} + \alpha_\mathrm{int})S + \beta\frac{N}{\tau}. \tag{B1.2.35}$$

Equation (B1.2.34) gives the time dependence of the carrier density. The first term on the right-hand side is the pump term, where I is the current supplied to the laser, e the unit charge and V the active volume of the laser. The second term accounts for spontaneous recombination with τ being the spontaneous lifetime. Finally, the last term accounts for stimulated recombination with Γg_act being the modal gain. Here Γ is the confinement factor, g_act the gain in the active region (cf section B1.2.2), and v_g the group velocity of the light in the laser. The group velocity appears because the rate equations describe the development of the carrier and photon densities in time, whereas the gain (and loss) describes the spatial evolution of the photon density.

The second rate equation (B1.2.35) describes the time dependence of the photon density. The first term on the right-hand side is recognized as the stimulated recombination term. The second term accounts for losses, with α_end describing facet losses and α_int being the internal loss coefficient. The final term occurs because a fraction β of the spontaneous emission events adds a photon to the lasing mode.

Under static conditions, the left-hand sides of equations (B1.2.34) and (B1.2.35) vanish, and if the current is below the threshold, the photon density is very small. Equation (B1.2.34) then reduces to the relation between current and carrier density, as given by equation (B1.2.8) in section B1.2.2. Under static conditions above the threshold, equation (B1.2.34) gives

$$S = \frac{I - I_\mathrm{th}}{eV}\frac{1}{v_\mathrm{g}\Gamma g_\mathrm{act}}. \tag{B1.2.36}$$

In order to translate the photon density S to optical power, we first multiply by the photon energy $h\nu$ to get the optical energy density, then by the active volume V to get the total optical energy in the laser, and finally by the facet loss rate $v_\mathrm{g}\alpha_\mathrm{end}$ to get the emitted power (from both ends of the laser). Hence

$$P = v_\mathrm{g}\alpha_\mathrm{end}Vh\nu S. \tag{B1.2.37}$$

Combining equations (B1.2.36) and (B1.2.37) and using the gain condition equation (B1.2.9) (or alternatively using equation (B1.2.35) for steady state with the last term neglected) gives the relation between optical power and current

$$P = (I - I_\mathrm{th})\frac{\alpha_\mathrm{end}}{\alpha_\mathrm{end} + \alpha_\mathrm{int}}\frac{h\nu}{e} \tag{B1.2.38}$$

in agreement with the results from section B1.2.2.

In writing the rate equations in the form given by equations (B1.2.34) and (B1.2.35), we have implicitly assumed that there was only one lasing mode. If several modes are lasing, there will be a photon density rate equation like equation (B1.2.35) for each mode, and the last term on the right-hand side of the carrier density rate equation, equation (B1.2.34), has to be summed over all lasing modes.

Before starting to derive results from the rate equations, some of the terms need to be discussed in more detail. Looking first at the second term on the right-hand side of equation (B1.2.34), we rewrite it as a polynomial expression

$$\frac{N}{\tau} = \frac{N}{\tau_{\mathrm{nr}}} + BN^2 + CN^3. \tag{B1.2.39}$$

In this expression, the first term accounts for *nonradiative recombination*, the second for *radiative recombination*, and the third term for *Auger recombination*. This latter is a higher order process which is particularly important at high carrier densities. It is seen that the carrier lifetime, τ, is not a constant, but depends on the carrier density.

Under conditions where the variations in the carrier density are small, it is convenient to introduce a *differential carrier lifetime*, defined by

$$\frac{1}{\tau'} = \frac{\mathrm{d}}{\mathrm{d}N}\left(\frac{N}{\tau}\right) = \frac{1}{\tau_{\mathrm{nr}}} + 2BN + 3CN^2. \tag{B1.2.40}$$

For the case where the three recombination term in equation (B1.2.39) are comparable, it follows that the differential carrier lifetime from equation (B1.2.40) is close to half the average carrier lifetime. Carrier lifetimes are typically in the range between one and a few nanoseconds.

In the last term in equation (B1.2.35) β is the fraction of the spontaneous emission events which add a photon to the lasing mode. Because the total number of modes is proportional to the laser volume V, it follows that β is proportional to the reciprocal of the volume. It is convenient to rewrite the spontaneous emission term as

$$\beta\frac{N}{\tau} = \frac{v_{\mathrm{g}}\Gamma g_{\mathrm{act}} n_{\mathrm{sp}}}{V} = \frac{R_{\mathrm{sp}}}{V} \tag{B1.2.41}$$

where n_{sp} is known as the *population inversion factor* and usually has a value between 1 and 2, with a value of 1 corresponding to complete inversion. Using equation (B1.2.41) in equation (B1.2.35), and comparing the first term with the third term, shows that we have VS photons due to stimulated emission in the laser and n_{sp} photons due to spontaneous emission. For operation close to, or above, the lasing threshold, the carrier density is close to or equal to the threshold carrier density, and we can replace Γg_{act} by g_{th} in equation (B1.2.41).

It turns out, that the rate equations in the form given by equations (B1.2.34) and (B1.2.35) do not agree fully with the observed laser behaviour. It is found that better agreement is obtained by including a nonlinear gain term, according to

$$g_{\mathrm{act}} \Rightarrow g_{\mathrm{act}}(1 - \varepsilon S) \tag{B1.2.42}$$

where ε is the *gain suppression parameter*. Such a nonlinear term may be somewhat surprising, because semiconductor lasers are usually regarded as being *homogeneously broadened*, i.e. a change in the carrier density (and hence the gain) due to stimulated recombination, will affect the gain over the whole spectral range. However, there is evidence that some degree of spectral hole burning (*inhomogeneous broadening*) is present, and carrier dynamics may also play a role.

The nonlinear gain gives rise to additional damping, and also explain why the spectrum is broader (i.e. contains more longitudinal modes) than might otherwise have been expected.

Because the rate equations are nonlinear, due to the dependence of the gain on the carrier density, they cannot in general be solved analytically, but a number of important results can still be derived from them. In the case of weak modulation, where the current I consists of a bias current plus a superimposed small signal modulation current, the rate equations can be linearized by expanding the current, the carrier density and the photon density around an operating point according to

$$X(t) = \overline{X} + \Delta X \exp(\,\mathrm{j}\omega t) \tag{B1.2.43}$$

with $X = I, N, S$ and $\Delta X \ll \overline{X}$. With a linear dependence of gain on carrier density, equation (B1.2.26), using equations (B1.2.40)–(B1.2.42), we eliminate the steady-state solution, and assume that the carrier density is close to the threshold carrier density. By discarding terms containing products of ΔX, it can be shown (see, for example [2] or [5] on the further reading list for details) that the laser modulation response has the form

$$\frac{\Delta S}{\Delta I} = \frac{1}{eVv_{\mathrm{g}}\Gamma g_{\mathrm{act}}} H(\omega), \quad H(\omega) = \frac{1}{1 + \dfrac{\mathrm{j}\omega\gamma}{\omega_{\mathrm{R}}^2} - \left(\dfrac{\omega}{\omega_{\mathrm{R}}}\right)^2}. \tag{B1.2.44}$$

The resonance frequency is given by

$$\omega_{\mathrm{R}} = \sqrt{v_{\mathrm{g}}^2 \Gamma a g_{\mathrm{th}} \overline{S}} \tag{B1.2.45}$$

where a is the gain slope from equation (B1.2.26), and g_{th} is the threshold gain from equation (B1.2.9). Because the photon density is proportional to the power, equation (B1.2.37), it follows that the resonance frequency increases with the square root of the power. It can be shown that the *damping rate* γ varies linearly with the square of the resonance frequency. The degree of damping is determined by spontaneous emission and by the nonlinear gain.

Resonance frequencies well in excess of 10 GHz can be achieved, and it is found that the resonance frequency is a reasonable measure of how fast the laser can be modulated. In the case of low damping, the 3 dB cut-off frequency is about 50% above the resonance frequency. For modulation frequencies above the resonance frequency, the laser response drops off rapidly. Ultimately, the 3 dB frequency will be limited by damping, because the damping increasing with increasing resonance frequency. Maximum modulation frequencies above 30 GHz have been reported. From equation (B1.2.45), it is seen that for achieving a high value of the resonance frequency, the laser should be operated high above threshold, and be designed to have a high value of the gain slope parameter.

Because the rate equations deal with the interaction of carriers and photons in the active region, other factors can influence the laser dynamics. *Parasitic elements*, such as the combination of a series resistance and a parallel capacitance, may reduce the achievable modulation frequency. *Carrier transport effects* can also lead to a roll off in the laser response at high frequencies. Other results that can be derived from the linearized rate equations include *harmonic distortion* and *intermodulation products* [3].

In general, an investigation of the large signal properties will require a numerical solution of the rate equations. An important practical case is the *turn on transient* when the current is changed from a value close to, or even below, threshold to an 'on' value well above threshold (see figure B1.2.10). The behaviour is explained as follows: Below threshold the photon density is very small, and the last term in the carrier density rate equation can be neglected. When the current is increased, the carrier density will increase and even exceed the threshold carrier density. As the gain is now above the lasing threshold, the photon density will increase rapidly, but when the photon density is sufficiently high, the carrier density will start to drop due to the stimulated emissions. When the carrier density drops back below the threshold carrier density, the photon density will start to fall, but when the photon density gets low, the carrier density will start to increase again. This process goes on until a steady-state situation is reached, with a photon density (and hence power) corresponding to the current. The frequency of the *ringing* in the turn on transient is equal to the resonance frequency corresponding to the 'on' level.

For digital modulation the *turn on* and *turn off* times are of importance. Analytical approximations for these times are given in [4].

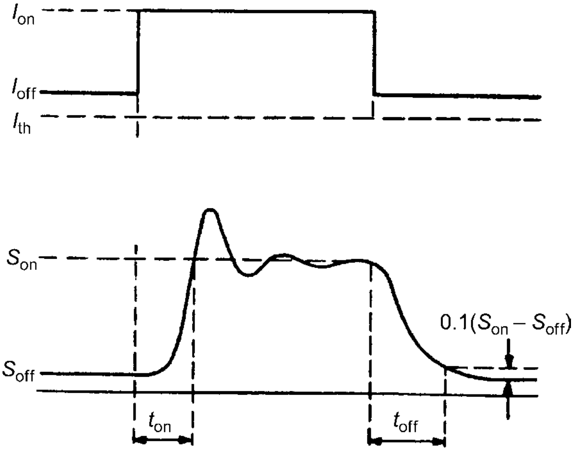


Figure B1.2.10. Digital modulation with the laser current being switched between 'off' and 'on' levels. The figure shows the turn on transient with the accompanying ringing, after [4].

$$t_{\mathrm{on}} = \frac{\sqrt{2}}{\omega_{\mathrm{R}}}\sqrt{\ln\left(\frac{S_{\mathrm{on}}}{S_{\mathrm{off}}}\right)} \qquad \text{(B1.2.46)}$$

$$t_{\mathrm{off}} = \frac{2\pi(\sqrt{2}-1)}{\omega_{\mathrm{R}}} + \frac{\varepsilon}{v_{\mathrm{g}}\Gamma a}. \qquad \text{(B1.2.47)}$$

The resonance frequency appearing in equations (B1.2.46) and (B1.2.47) is the one corresponding to the 'on' level. A careful balance must be found in setting the 'off' level. If it is too low, the turn on time will be long, and if it too high the *extinction ratio*, defined as the ratio between the 'on' and the 'off' levels will be too low.

Taking into account that spontaneous emissions do not occur at a constant rate, but are the result of a statistical process, it is possible to derive results for the laser *relative intensity noise* (RIN) of the laser power, and its spectral distribution. The frequency dependence of the RIN can be written as

$$\mathrm{RIN}(\omega) = |H(\omega)|^2 \frac{\left(\frac{1}{\tau'} + v_{\mathrm{g}}\Gamma a S\right)^2 + \omega^2}{\omega_R^4} \frac{2R_{\mathrm{sp}}}{VS} \qquad \text{(B1.2.48)}$$

where $\mathrm{RIN}(\omega)$ is the relative intensity noise per unit bandwidth around ω. The total relative intensity noise is found by integrating equation (B1.2.48) over the relevant bandwidth.

Readers should consult [5] on the further reading list for more details on modulation and noise properties of semiconductor lasers.

The *laser linewidth* (i.e. the spectral width of a single longitudinal mode) is also related to spontaneous emission. Photons added to the lasing mode by spontaneous emission events have a random phase relative to the optical field in the laser. This gives rise to phase fluctuations which can be quantified in terms of a finite laser linewidth. The laser linewidth can be shown to be proportional to the spontaneous emission rate R_{sp} and inversely proportional to the photon density. Typical linewidth values for a single frequency laser are a few tens of megahertz, but much lower values have been reported for optimized structures.

B1.2.7 Visible wavelength lasers

For many years, nearly all semiconductor lasers were based on GaAs substrates. This is convenient for forming heterostructures using $Al_xGa_{1-x}As$ compounds, with a relatively low Al content (if any) in the active layer, and a higher Al content in the confinement layers. In order to provide sufficient confinement of carriers, the Al fraction in the confinement layers must exceed that of the active region by about 20–30%, but, on the other hand, Al fractions above 40% cannot be used because these materials do not have a direct bandgap.

By varying the Al content in the active layer, lasing wavelengths can be achieved ranging from 880 nm (with no Al in the active layer) to about 750 nm. These wavelengths are just beyond the visible part of the spectrum, in the near infrared. Lasers operating around 780 nm are used extensively in CD players and CD-ROM drives.

For use in storage devices, it is an obvious advantage to operate at short wavelengths. Because the minimum features that can be resolved optically have a linear dimension of the order of the wavelength used, it follows that the recording density, which depends on the reciprocal of the focused spot area, is proportional to the inverse square of the wavelength. A reduction of the wavelength from 780 to 650 nm (used in DVD players) will increase the recording density by nearly 50%, and even higher densities will be possible by further reduction of the wavelength.

A possible material system is $(Al_xGa_{1-x})_yIn_{1-y}P$. A lasing wavelength in the 670–680 nm region is achieved using GaAs substrates, a GaInP active region, and having $y = 0.5$ and $x \approx 0.7$ in the confinement layers. Shorter wavelengths are possible by using quantum well (possibly strained) active regions.

Wavelengths in the 600–700 nm region lie within the visible part of the spectrum, and lasers operating in this range are of interest for applications such as bar code scanners and pointers, where they can replace HeNe lasers, which operate at 633 nm.

Even shorter wavelengths (i.e. green and blue) are possible, and currently there is a great deal of activity in developing lasers in the $(Al_xGa_{1-x})_yIn_{1-y}N$ system, based on GaN substrates. Blue emitting LEDs are already widely available. A particular problem with the nitride materials is that their crystal structure makes it more difficult to form good end facets by cleaving.

An interesting aspect of the development of short wavelength semiconductor lasers is that it gives access to the whole visible range, thus, for example, allowing fabrication of high brightness displays.

B1.2.8 GaAs-based VCSELs

In a *vertical cavity surface emitting laser* (VCSEL), the direction of lasing is perpendicular to the active layer, and light is emitted from the top or bottom surface of the chip. As this means that the active cavity is very short, it follows from equation (B1.2.6) that very high end-reflectivities are required. Such high reflectivities can be achieved by producing a stack consisting of a large number of layers of alternating high and low refractive index, each layer having a thickness corresponding to a quarter of a wavelength.

The *field reflection* for an interface between a layer of index n_1 and one of index n_2 is given by the *Fresnel formula*

$$r = \frac{n_1 - n_2}{n_1 + n_2} = \frac{\Delta n}{2n_{\mathrm{av}}}. \tag{B1.2.49}$$

The *power reflection* is given by the square of the field reflection. With each layer being a quarter of a wavelength in thickness, we introduce *the coupling coefficient*, κ, as the reflection per unit length (note that, by a quarter of a wavelength, we mean a quarter of a wavelength in the material)

$$\kappa = \frac{2\Delta n}{\lambda}. \tag{B1.2.50}$$

In a stack with N layer pairs, i.e. $2N$ layers, the total reflection (not including the final surface/air or surface/substrate interface) can be shown to be given by

$$r_{\mathrm{tot}} = \tanh\left(2N\frac{\Delta n}{2n_{\mathrm{av}}}\right) = \tanh(\kappa t) \tag{B1.2.51}$$

where the total thickness of the reflector stack is

$$t = 2N\frac{\lambda}{4n_{\mathrm{av}}}. \tag{B1.2.52}$$

See, for example [5] for more details. As an example, we consider a stack with 30 layer pairs, an index difference of 0.3 and an average index of 3.3. The power reflectivity, given as the square of the expression in equation (B1.2.51) is in this case 0.98.

The *reflection bandwidth* of a reflector stack depends on the wavelength and on the *index contrast*

$$\Delta\Lambda = \frac{\lambda}{2}\frac{\Delta n}{n_{\mathrm{av}}}. \tag{B1.2.53}$$

It is also useful to define an *effective length* of the stack, as a measure for how far the light 'on average' penetrates into the stack. This quantity is given by

$$L_{\mathrm{eff}} = \frac{1}{2\kappa}\tanh(\kappa t). \tag{B1.2.54}$$

For $\kappa t > 1$ we then have

$$L_{\mathrm{eff}} \approx \frac{\lambda}{4\Delta n}. \tag{B1.2.55}$$

Including both a top and a bottom reflector stack, we see that the effective cavity length of a VCSEL is much smaller than that of a conventional edge emitting laser, typically by a factor of between 10 and 100. Combined with the fact that in a VCSEL there is only gain in a part of the cavity, it follows that reflectivities of the order of 0.99 are necessary. A schematic of a VCSEL is shown in figure B1.2.11.

It is a considerable advantage of the short cavity of VCSELs that because of the resulting wide mode spacing, only one longitudinal mode has significant gain, cf equations (B1.2.2)–(B1.2.3). This means that a VCSEL is by its nature a single frequency laser. Other major advantages include: the possibility of matching the laser spot size to that of a fibre—making coupling simpler and more efficient, and the use of on-wafer testing in the fabrication process. See [6] for more details and a review.

The various advantages of GaAs VCSELs make them highly suitable as relatively low cost transmitters for short distance systems operating at relatively short wavelengths, such as data links. The technology for VCSELs based on InP and intended for operation at the 'telecoms' wavelengths of 1300 and 1550 nm has proved to be considerably more difficult.

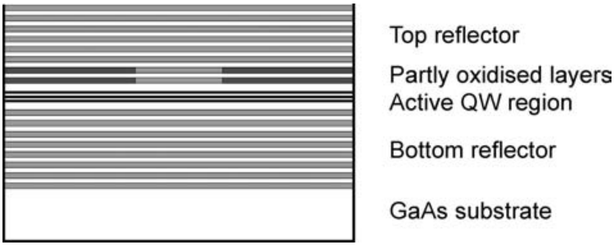


Figure B1.2.11. Simplified diagram of a VCSEL on GaAs substrate. In this case, the active material is based on InGaAs quantum wells (see section B1.2.9). The top and bottom reflectors consist of alternating layers of GaAs and AlAs. Some of the top AlAs layers have been partly oxidized, forming low index, nonconducting Al_xO_y, which gives both optical and electrical confinement. The actual number of layers in the reflector stacks is higher than shown.

For long wavelength materials, the index contrast that can be achieved in the mirror stacks is smaller than for shorter wavelength materials, hence more layers are needed. In addition, the electrical and thermal properties of the long wavelength materials tend to be inferior to those for the short wavelength material. In addition, higher order recombination processes, cf equation (B1.2.39) present more of a problem.

One possible way of overcoming some of these problems is the use of, for example, 980 nm lasers for optical pumping, as an alternative to direct electrical pumping. Other options for long wavelength VCSELs include the use of hybrid structures with dielectric (nonsemiconductor) reflectors, or the use of new semiconductor material combinations.

B1.2.9 High power pump lasers

In general, semiconductor lasers have very high efficiencies, i.e. a large fraction of the electrical input power supplied to the laser is converted to laser power. The overall efficiency, also known as the *wallplug efficiency* (defined as the ratio between the optical output power and the electrical power supplied to the laser), can exceed 50%. The combination of the high efficiency with the ability to design lasers to operate within a given spectral range, makes the semiconductor laser very attractive as a pump source for other lasers, because a high optical power can be delivered at a wavelength coinciding with a pump band in the other laser.

In section B1.2.2, we assumed that the two laser facets had the same reflectivity, R. In this case, equal amounts of power will be emitted from the two ends. This power balance can be changed by changing one or both end reflectivities, e.g. by coating. Equation (B1.2.6) in section B1.2.2 and equations (B1.2.37)–(B1.2.38) in section B1.2.6 still remain valid, provided we interpret R as the geometric mean of the two end reflectivities

$$R = \sqrt{R_{\text{front}} R_{\text{back}}}. \tag{B1.2.56}$$

It is easy to show that the power ratio is given by

$$\frac{P_{\text{front}}}{P_{\text{back}}} = \sqrt{\frac{R_{\text{back}}}{R_{\text{front}}}} \frac{1 - R_{\text{front}}}{1 - R_{\text{back}}}. \tag{B1.2.57}$$

By having a high back reflectivity and a low front reflectivity most of the power is emitted from the front. The facet reflectivities are subject to a design trade-off: A high value of R will give a low end loss, equation (B1.2.6), leading to a low threshold gain, equation (B1.2.9), and hence a low threshold current; but the differential efficiency, equation (B1.2.10), will be low. On the other hand, a low value of R will give a high efficiency, but also a high threshold current.

The limiting factor for high power operation is optical damage to the end facet. This occurs when the optical power density gets too high. By using proper facet coating technology, the power density can be of the order of $10\,\mathrm{MW\,cm^{-2}}$ before damage occurs. This implies a design trade-off for the waveguide in the laser: In order to have a well-defined beam, the waveguiding will have to be reasonably strong, but a strong waveguide will tend to give a small spot size, and may also make it possible to excite higher order modes. Single transverse mode, high power, lasers usually have spot sizes of a few square micrometres, giving power levels up to a few hundred milliwatts.

A very important practical example of the use of semiconductor lasers as a pump source is that intended to power the *erbium-doped fibre amplifier* (EDFA), which is used to provide amplification for wavelengths in the 1530–1560 nm range. The EDFA can be pumped at several wavelengths, the most commonly used being 980 nm (1480 nm is another possible pump wavelength).

Using conventional materials on GaAs or InP substrates, there are no material combinations which can be used to form a laser operating at 980 nm. However, strained InGaAs has a narrower bandgap than GaAs, so a possible structure consists of one or more strained $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ quantum well(s) (about 20% In, and a well width around 10 nm) in a separate confinement structure, cf figure B1.2.9, with a pair of GaAs layers providing carrier confinement, and with $\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As}$ layers for optical confinement.

It is a characteristic feature of semiconductors, that the bandgap energy (and hence the wavelength region of gain) varies with temperature. If the lasing wavelength is not controlled a drift of about 0.4–0.5 nm per degree will occur in the peak wavelength of the gain, and hence of the laser. Consequently, lasing would only correspond to the EDFA pump band for a limited temperature range. In addition, fabrication tolerances on layer compositions and thicknesses give some variation in the lasing wavelength from device to device. These effects can be mitigated by using an external reflector with spectrally well-defined (and temperature stable) reflection properties. An example of a suitable reflector is a *fibre Bragg grating* (FBG) with a reflection of a few per cent, comparable to the reflection of the (anti-reflection coated) laser facet, and a reflection bandwidth of the order of 1 nm.

Semiconductor lasers are also widely used for pumping of solid-state crystal-host lasers, e.g. the Nd:YAG laser that has a pump band near 810 nm. Extremely high power levels (over 100 W) can be achieved in structures with a wide active region, or with multiple active stripes [7]. By forming a stack of such devices, power levels of as much as 1 kW for a $1\,\mathrm{cm}^2$ emitting area have been obtained.

B1.2.10 Fixed wavelength DFB lasers

The spectral width of the gain curve of semiconductor lasers, cf figure B1.2.7, exceeds the modespacing, given by equation (B1.2.3), by a large factor, leading to possible multimode lasing. For applications in fibre optical communication, this gives two problems: Firstly, the resulting spectral width of several nanometres leads to pulse broadening due to fibre dispersion, and secondly, the number of distinct channels that can be multiplexed on the same optical fibre is reduced.

In order to achieve single mode lasing, some form of spectral selectivity is necessary. The conventional way to achieve this is by incorporating a *periodic structure* (i.e. a *grating*) in the laser, as shown schematically in figure B1.2.12. In this structure, the grating provides internal reflections at a wavelength determined by the grating period. This type of laser is known as a *distributed feedback laser* or *DFB* in short.

The wavelength selected by the grating is given by

$$\lambda_{\mathrm{DFB}} = 2n_{\mathrm{eff}}\Lambda \tag{B1.2.58}$$

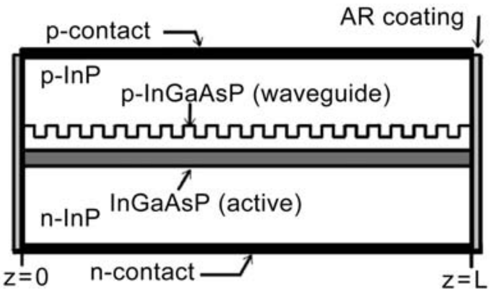


Figure B1.2.12. Outline of a DFB laser. The waveguide layer has a bandgap (and consequently a refractive index) between that of the active layer and InP.

where n_{eff} is the effective index and Λ the grating period. Because the grating only provides efficient internal feedback at wavelengths very close to λ_{DFB}, any wavelength different from λ_{DFB} will have a higher rate of loss through the end facets, and as a result lasing will occur at λ_{DFB}.

As the grating period, Λ in equation (B1.2.58), is equal to half a wavelength, the square grating shown in figure B1.2.12 is in fact similar to the quarter wave stack considered in section B1.2.8. However, in contrast to the situation shown in figure B1.2.11, it is only a part of the optical field that 'sees' the region where the refractive index varies. In other words, it is the *effective* (modal) refractive index that varies, and from perturbation theory, we find this variation to be

$$\Delta n_{\mathrm{eff}} = \Gamma_{\mathrm{g}}(n_1 - n_2) \tag{B1.2.59}$$

where Γ_{g} is the confinement factor for the grating (the ratio of the optical power in the grating relative to the total optical power) and n_1 and n_2 the refractive indices of the materials forming the grating.

In analogy with equation (B1.2.50) the coupling coefficient is

$$\kappa = \frac{2\Delta n_{\mathrm{eff}}}{\lambda}. \tag{B1.2.60}$$

The bandwidth of the grating is found from equation (B1.2.53), but with Δn replaced by Δn_{eff}.

As an example, we consider a wavelength of 1.55 µm, and an average effective refractive index of 3.3. For $(n_1 - n_2) = 0.1$ and $\Gamma_{\mathrm{g}} = 0.1$, we find $\kappa = 130\,\mathrm{cm}^{-1}$, a length of 200 μm gives $\kappa L = 2.6$, and the bandwidth is 2.3 nm. As seen in this example, the bandwidth is similar to the modespacing, so only a single mode will be selected by the grating.

The κL product in a DFB laser plays a role similar to the facet reflectivity in an ordinary laser, and a high value (i.e. strong feedback) will give a low threshold current, but also a low efficiency. As the lasing mode experiences gain from the active layer and, at the same time, feedback from the grating, the gain condition and the phase condition (equivalent to equations (B1.2.9) and (B1.2.2), respectively) are coupled in a DFB laser, and the wavelengths of the modes and their required gains are found by solving a complex transcendental equation.

The discrete (and nonselective) reflections from the end facets will interfere with the distributed reflection from the grating, and usually the reflection from the front facet is suppressed by applying an *anti-reflection* (AR) coating. The detailed theory for DFB lasers is rather involved, particularly if the facet reflections are included, and readers are referred to [6, 7] on the further reading list for more information.

Gratings can be fabricated by covering the waveguide layer with a photoresist, which is then exposed to an optical interference pattern, and the developed resist is used as an etch mask. After etching of the grating, the remaining layers in the laser structure are grown.

As mentioned in section B1.2.9, the bandgap energy of a semiconductor is temperature dependent, and the wavelength corresponding to the peak gain, will change by about 0.4–0.5 nm per degree. However, in a DFB laser the lasing wavelength is stabilized and determined by the grating rather than by the gain curve, so a smaller temperature coefficient due to refractive index change will apply. The refractive index of a semiconductor laser structure is still quite sensitive to temperature; so this will lead to a temperature dependence of the lasing wavelength. A typical value is about 0.1 nm per degree (corresponding to a change in the optical frequency of about 10 GHz per degree).

Whereas temperature tuning can be used to trim the wavelength to a given value, the temperature dependence also means that, in order to ensure that the lasing frequency is within 10 GHz of a given value, the laser temperature has to be stabilized to within 1 degree.

DFB lasers are often designed for direct modulation at data rates up to 2.5 Gbit/s. For higher data rates external modulation is used. An interesting option is to integrate, monolithically, a DFB laser and an *electroabsorption modulator* (EAM).

In systems using *wavelength division multiplexing* (WDM), signals from several lasers are multiplexed before transmission, and in order to ensure interoperability of equipment from different manufacturers, the International Telecommunications Union, ITU, has set a standard for optical transmission frequencies. This standard is based on a frequency grid with a 100 GHz spacing. Consequently, the range from 192.1 THz (corresponding to a wavelength of 1560.61 nm) to 195.9 THz (1530.33 nm) comprises 39 channels.

B1.2.11 Tunable lasers

The wide optical gain curve in a semiconductor laser makes it possible to achieve tuning of the lasing wavelength/frequency. As already mentioned, tuning is possible by changing the operating temperature. However, unless a large temperature variation is allowed, the tuning range will be limited to a few nanometres in wavelength (a few hundred gigahertz in frequency), and thermal tuning is comparatively slow (microsecond to millisecond range).

The fact that the refractive index depends on the carrier density can be used for tuning. However, in a simple structure (like figure B1.2.3 or B1.2.4), and also in DFB lasers, the carrier density is clamped to the value which is required to give sufficient gain to satisfy the lasing condition, and tuning by carrier density changes is not possible. This limitation can be overcome by using structures with two (or more) separate regions. One example is the *distributed Bragg reflector* (DBR), laser shown in figure B1.2.13. The tuning speed will be limited by the carrier lifetime in the tuning region (nanosecond range).

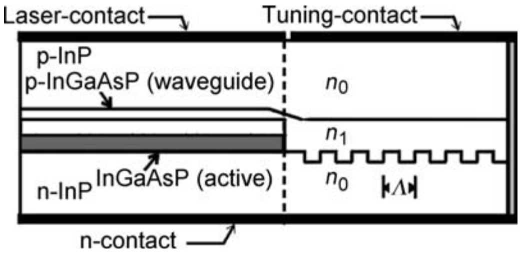


Figure B1.2.13. Two section DBR laser. The output power is controlled by the laser current supplied to the active region. The tuning current supplied to the Bragg reflector region controls the carrier density in that region, and hence its refractive index. This in turn tunes the wavelength at which the grating gives efficient reflection. Tuning ranges can be up to 10–15 nm, e.g. [8].

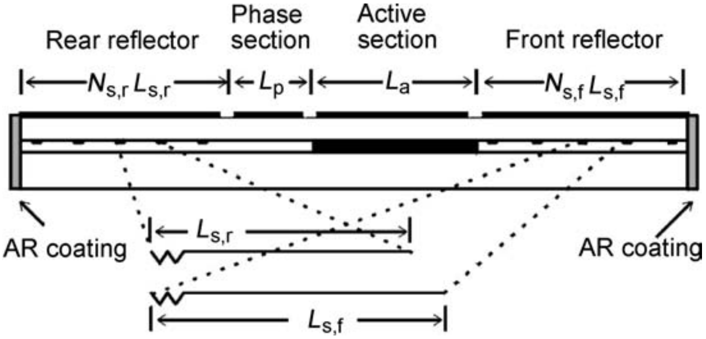


Figure B1.2.14. Sampled grating DBR. Instead of a continuous grating the two reflectors have sampled gratings. These gratings give reflection spectra that have a comb of reflection peaks with a spacing determined by the sampling period. By using two different sampling periods, two reflection combs with different periodicities are obtained.

It should be noted that the tuning of a DBR laser is not continuous, but shows jumps between the wavelengths that satisfy the resonance condition given by equation (B1.2.2).

In order to give separate control of the lasing wavelength, a phase section can be inserted between the active section and the DBR reflector. This section is similar to the DBR section, but it has no grating. Changing the carrier density, and hence the refractive index in this section, makes it possible to fine tune the lasing wavelength.

Whereas the tuning range of a two-section DBR laser is limited by how much the refractive index of the tuning section can be changed, wider tuning ranges can be achieved using somewhat more complicated structures. An example of such a structure is the *sampled grating DBR* (SGDBR) which is shown in figure B1.2.14.

A small change of the refractive index of one of the tuning sections gives a large change in wavelength, because a new pair of reflection peaks will coincide. This behaviour is recognized as the *Vernier effect*, see [9] for more details. This principle leads to greatly enhanced tuning ranges, with up to about 100 nm being reported. Similar tuning ranges have also been achieved by combining a tunable co-directional coupler with a sampled grating [10].

A particular problem with the tunable laser structures described earlier is that several control currents are required to set a specific combination of power and wavelength. Tunable lasers must therefore be characterized in sufficient detail to identify the current combinations required for various wavelengths. The laser driver electronics must contain this information in such a form that tuning can be achieved in response to simple external instructions.

Other semiconductor laser structures capable of very wide tuning include various forms of external cavity lasers. Ultimately, the tuning range is of course limited by the width of the gain curve. More details on different tunable laser types can be found in reference [8] on the further reading list.

Figure B1.2.15 shows an example of a widely tunable laser.

B1.2.12 Quantum cascade lasers

The *quantum cascade laser* is radically different from other semiconductor lasers. The lasing is based on inter-subband transitions, and involves only one type of carrier (the electrons); it is therefore also referred to as a *unipolar* laser.

The active region consists of a number of coupled quantum wells, and the possible energy states form a set of *minibands* and *minigaps*. An electrical field is applied to the structure and the electrons enter

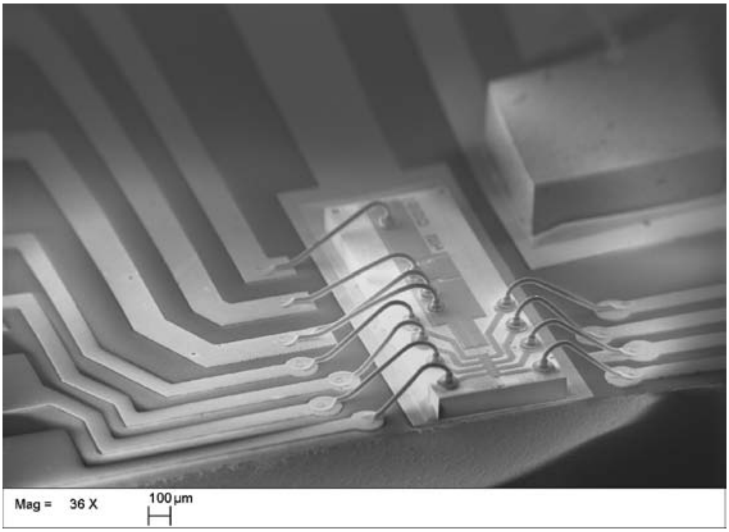


Figure B1.2.15. Micrograph of widely tunable laser from Bookham Technology. This particular structure, known as the 'digital supermode structure' has multiple individually contacted gratings in the front section. Courtesy Bookham Technology.

the active region from an *injector* region with aligned energy levels, creating a population inversion. A schematic diagram is shown in figure B1.2.16.

The successful fabrication of a quantum cascade laser requires a detailed design of the quantum well and barrier thicknesses, the material compositions and the doping levels, as well as a high level of

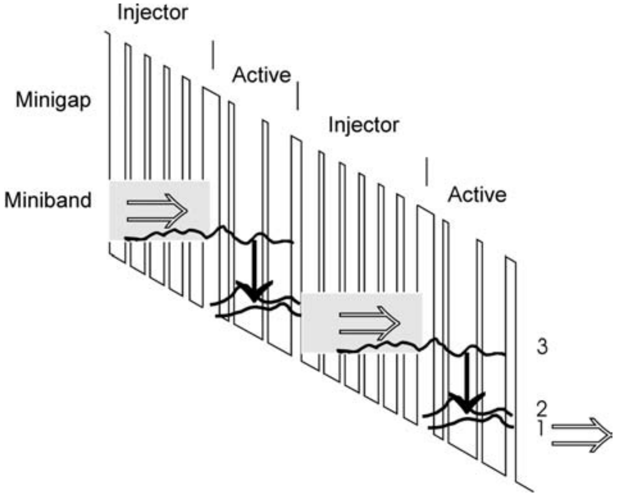


Figure B1.2.16. Basic principle of a quantum cascade laser. The figure shows the band edge for the conduction band subjected to an electrical field. Some of the electron wavefunctions are indicated, each with a baseline corresponding to the relevant energy level. As shown, the laser structure can have a number of stages, and by regarding the miniband in the injector as a pump band, the structure looks like a set of four level lasers in series. The total thickness of one injector region and one active region is about 50 nm, based on reference [11].

accuracy in the materials growth (i.e. use of MBE). A possible combination of materials is, for example, GaInAs for the quantum wells and AlInAs for the barriers.

Quantum cascade lasers have been made for operation at wavelengths ranging from a few micrometres to above 20 μm. Usually, they are either operated under pulsed conditions or at a low temperature, but examples of continuous operation at room temperature have also been reported. Whereas these lasers usually operate in several longitudinal modes, the spectral properties can be improved by the use of a grating to form a DFB laser (see section B1.2.10). Reference [11] contains a recent review of quantum cascade lasers.

B1.2.13 Concluding remarks

The main objectives of this chapter have been firstly to provide an introduction to the basic concepts related to semiconductor lasers, and secondly to show examples of semiconductor laser structures optimized for various functions and applications. In doing so, a number of details and other aspects have necessarily been left out, and the reader is encouraged to consult the further reading list for additional information. We note that somewhere around 40 books dealing exclusively or mainly with semiconductor lasers have been published to date, and there are whole books devoted to topics that have been covered on a single page in this chapter. In addition, some topics have not been covered at all in this chapter, including, for example, the use of modelocked lasers for short pulse generation, and the monolithic integration of lasers structures performing other functions such as modulation, detection, wavelength conversion, etc. Anyway, it is the hope of the author that this paper will serve a dual purpose as a basic reference, as well as a starting point for further study.

Acknowledgments

Some of the sections in this chapter are partly based on material used in the article: J Buus, ‘Optical Sources’ in *Wiley Encyclopedia of Telecommunications*, Editor J Proakis, to appear in 2002.

References

[1] *IEEE J. Quantum Electron.* **23**, Special Issue on Semiconductor Lasers, June 1987
[2] Willner AE ed, *IEEE J. Selected Topics Quantum Electron.* **6**, Millennium Issue, November–December 2000
[3] Darcie T E, Tucker R S and Sullivan G J 1985 Intermodulation and harmonic distortion in InGaAsP lasers *Electron. Lett.* **21** 665–666
[4] Tucker R S 1984 Large-signal switching transients in index-guided semiconductor lasers *Electron. Lett.* **20** 802–803
[5] Corzine S W, Yan R H and Coldren L A 1991 A tanh substitution technique for the analysis of abrupt and graded interface multilayer dielectric stacks *IEEE J. Quantum Electron.* **27** 2086–2090
[6] Iga K 2000 Surface-emitting laser—its birth and generation of new optoelectronics field *IEEE J. Selected Topics Quantum Electron.* **6** 1201–1215
[7] Sakamoto M, Endriz J G and Scifres D R 1992 120 W CW output power from monolithic AlGaAs (800 nm) laser diode array mounted on diamond heatsink *Electron. Lett.* **28** 197–199
[8] Delorme F, Grosmaire S, Gloukhian A and Ougazzaden A 1997 High power operation of widely tunable 1.55 μm distributed Bragg reflector laser *Electron. Lett.* **33** 210–211
[9] Jayaraman V, Chuang Z-M and Coldren L A 1993 Theory, design, and performance of extended tuning range semiconductor lasers with sampled gratings *IEEE J. Quantum Electron.* **29** 1824–1834
[10] Rigole P-J, Nilsson S, Bäckbom L, Klinga T, Wallin J, Stålnacka B, Berglind E and Stolz B 1995 114 nm wavelength tuning range of a vertical grating assisted codirectional coupler laser with a super structure grating distributed Bragg reflector *IEEE Photonics Technol. Lett.* **7** 697–699
[11] Capasso F, Gmachl C, Paiella R, Tredicucci A, Hutchinson A L, Sivco D L, Baillargeon J N, Cho A Y and Liu H C 2000 New frontiers in quantum cascade lasers and applications *IEEE J. Selected Topics Quantum Electron.* **6** 931–947

Further reading

Chuang S L 1995 *Physics of Optoelectronic Devices* (Chichester, UK: Wiley)

Coldren L A and Corzine S W 1995 *Diode Lasers and Photonic Integrated circuits* (Chichester, UK: Wiley)

Agrawal G P and Dutta N K 1993 *Semiconductor Lasers* (New York, NY: Van Nostrand Reinhold)

Zory P S ed, 1993 *Quantum Well Lasers* (Boston, MA: Academic Press)

Petermann K 1988 *Laser Diode Modulation and Noise* (Dordrecht, The Netherlands: Kluwer Academic)

Morthier G and Wankwikelberge P 1997 *Handbook of Distributed Feedback Laser Diodes* (Noorwood, MA: Artech House)

Carroll J E, Whiteaway J E A and Plumb R G S 1998 *Distributed Feedback Semiconductor Lasers* (Stevenage, UK: IEE)

Amann M-C and Buus J 1998 *Tunable Laser Diodes* (Noorwood, MA: Artech House)

B2

Optical detectors and receivers

Hidehiro Kume

B2.1 Introduction

This chapter describes the operating principles, construction and characteristics of major optical detectors.

Optical detectors (often called photodetectors) can be classified by their principle of light detection as shown in table B2.1. Typical optical detectors utilizing physical or chemical changes are photographic films, but these are now seldom used in photometric applications. Methods utilizing solid or gas ionization and scintillation are usually limited to detection in the x-ray and gamma-ray regions.

Optical detectors making use of photoelectric effects are widely used as UV to IR sensors in various applications including measurement instruments, industrial production equipment and communication devices. To make the contents of this chapter more practical, we will chiefly discuss optical detectors utilizing the photoelectric effects.

Optical detectors utilizing photoelectric effects can be further divided by their detection principle into two groups: one using external photoelectric effects and the other using internal photoelectric effects. Table B2.2 shows typical optical detectors utilizing these effects and their features. As can be seen from table B2.2, optical detectors also fall under two categories: point detectors that merely detect light intensity and two-dimensional detectors including position sensors and image sensors. Spectral response range (detectable wavelength band) also differs depending on the type of optical detector. In general, optical detectors utilizing the external photoelectric effect are represented by photomultiplier tubes (PMTs) and exhibit fast time response and high sensitivity. On the other hand, optical detectors using the internal photoelectric effect, such as photodiodes (PDs) and photoconductive cells, offer a wide spectral response range, compact size and easy operation. Along with recent trends towards image measurements, the product quality and quantity of two-dimensional detectors are improving and increasing.

B2.2 Photoelectric effects

As stated, photoelectric conversion is roughly divided into external photoelectric effects [1] by which bound electrons inside a semiconductor thin film are released into a vacuum when light strikes the semiconductor and internal photoelectric effects [2] where photoelectrons are generated inside a semiconductor by light and excited into the conduction band. The photocathode used as the photoemissive surface of a PMT has the former function. The photoconductive effect and photovoltaic effect take place by the latter principle.

Table B2.1. Classification of optical detectors.

Principle	Method	Detectable electromagnetic radiation
Physical or chemical changes	Dosimeters	Charged particles
	Photographic films	UV, visible, IR radiation, charged particles
	Cloud chambers, bubble chambers	X-rays, gamma-rays, charged particles
Solid or gas ionization	Proportional counters, etc	Charged particles, x-rays, gamma-rays
Scintillation	Solid state or liquid scintillators + optical detectors	Charged particles, x-rays, gamma-rays
Photoelectric effect	Optical detectors utilizing external photoelectric effect (photomultiplier tube, etc)	UV, visible, near IR radiation
	Optical detectors utilizing internal photoelectric effect (photodiodes, etc)	UV, visible, near IR radiation, x-rays, gamma-rays

Table B2.2. Types and classification of optical detectors.

Detecting principle		Detectors (point detectors)	Two-dimensional detectors	Spectral response range
External photoelectric effect		Phototubes Photomultiplier tubes	Image intensifiers Streak cameras	Vaccum UV to near IR
Internal photoelectric effect	Photoconductive effect	Photoconductive cells	Photoconductive type camera tubes (vidicons, etc)	Visible to IR
	Photovoltaic effect	Photodiodes Phototransistors Avalanche photodiodes	Semiconductor image sensors (CCD, etc) Semiconductor position sensors (PSD, etc)	UV to near IR
Thermal effect	Pyroelectric effect	Pyroelectric IR detectors	Pyroelectric image sensors	Near IR to far IR
	Photovoltaic type	Thermocouples		
	Conductivity type	Bolometers Thermistors		

B2.2.1 External photoelectric effect

Semiconductor thin films having a photoemissive surface are usually called a ‘photocathode’ [1, 3] and have a band model structure like that shown in figure B2.1. Inside a semiconductor, there exists a valence band occupied by electrons, a forbidden band that cannot be occupied by electrons and a conduction

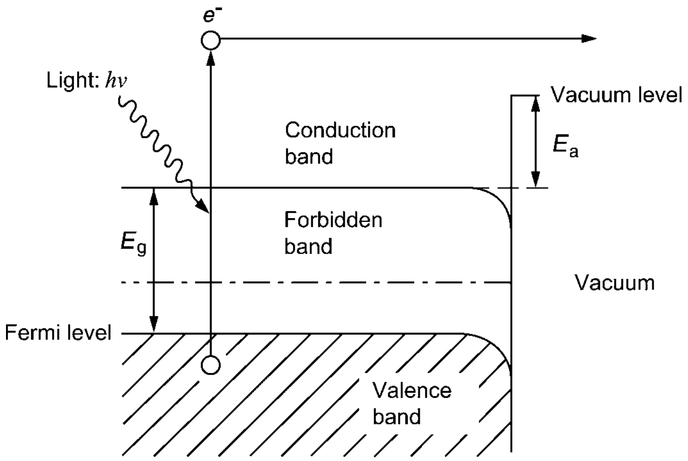


Figure B2.1. Photocathode band model.

band where electrons are free to move. When photons strike a photocathode, electrons in the valence band absorb photon energy *hv*, become excited into the conduction band and diffuse toward the photocathode surface. When the diffused electrons have enough energy to overcome the vacuum level barrier, they are emitted into the vacuum as photoelectrons. Photocathodes can be classified by photoelectron emission process into a reflection mode and a transmission mode. The reflection mode photocathode is usually formed on a metal plate and photoelectrons are emitted from the photocathode in the opposite direction to the incident light. The transmission mode photocathode is usually deposited as a thin film on an optically transparent flat plate and photoelectrons are emitted in the same direction as that of the incident light. Most photocathodes are made of a compound semiconductor chiefly consisting of alkali metals with a low work function such as Cs_3Sb, Na_2KSb, etc. Changing the photocathode materials makes it possible to achieve sensitivity in various bands of the spectrum from soft x-rays to near IR radiation (described later in detail).

PMTs [3] are the typical optical detectors utilizing the external photoelectric effect. They are vacuum tubes with a glass envelope and consist of a photoemissive cathode (photocathode), an electron multiplier and an electron collector (anode) in a vacuum tube. Figure B2.2 shows the schematic construction of a PMT. Light which enters a PMT is detected and produces an output signal through the following processes.

(1) Light passes through the input window and enters the photocathode in a vacuum.

(2) Excites the electrons in the photocathode so that photoelectrons are emitted into the vacuum (external photoelectric effect).

(3) Photoelectrons are accelerated and focused by the focusing electrode onto the first electrode called a ‘dynode’ in the electron multiplier section where they are multiplied by means of secondary electron emission. This secondary emission is repeated at each of the successive dynodes. (A dynode is an electrode capable of emitting secondary electrons. An electron multiplier used in a PMT usually has about 10 stages of dynodes.)

(4) A bunch of multiplied secondary electrons emitted from the last dynode are ultimately collected by the anode and output to an external circuit as an electrical signal. Since PMTs are a kind of

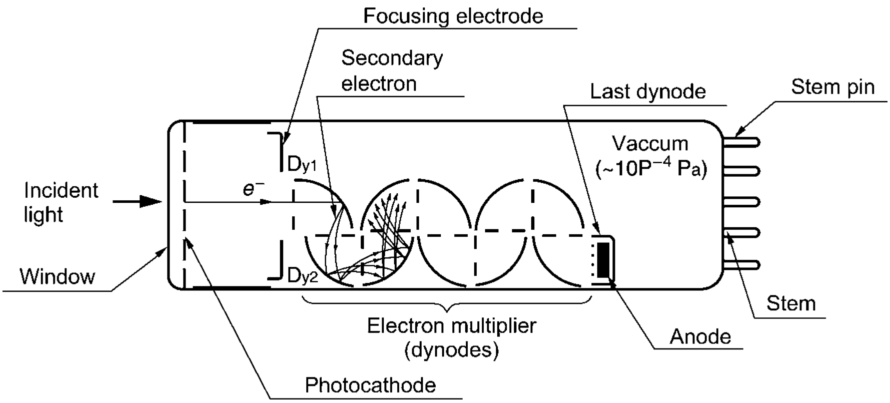


Figure B2.2. Construction of a photomultiplier tube.

electron tube, they are relatively large in size, but are superior in sensitivity and response speed, making them useful as optical detectors in a variety of applications such as analytical instruments, medical equipment and industrial measurement systems.

B2.2.2 Photovoltaic effect

Internal photoelectric effects will be explained by using a PD as an example.

A cross section of a typical PD is shown in figure B2.3(*a*). The p-layer at the light sensitive surface and the n-layer at the substrate form a p–n junction that serves as a photoelectric converter. The usual p-layer for a silicon PD is formed by selective diffusion of boron to a thickness of approximately 1 μm and the neutral region at the junction between the p-layer and n-layer is known as the depletion layer. By varying and controlling the thickness of the outer p-layer, substrate n-layer and bottom N+ layer as well as the doping concentration, the PD's spectral response and frequency response can be controlled.

A band model for photoelectric conversion that occurs at the p–n junction of a PD is shown in figure B2.3(*b*). When light strikes a PD and the light energy is greater than the band gap energy E_g, the electrons in the valence band are excited and pulled up into the conduction band, leaving holes in their place in the valence band. These electron–hole pairs are generated throughout the p-layer, depletion layer and n-layer. In the p-layer and depletion layer the electric field accelerates the electrons toward the n-layer and the holes toward the p-layer. Of the electron–hole pairs generated in the n-layer, the electrons are left in the n-layer conduction band along with electrons that have arrived from the p-layer, while the holes diffuse through the n-layer up to the p–n junction and collect in the p-layer valence band while being accelerated. In this manner, electron–hole pairs generated in proportion to the amount of incident light are accumulated in the n-layer and p-layer, resulting in a positive charge in the p-layer and a negative charge in the n-layer. When an external circuit is connected between the p-layer and n-layer, electrons will flow from the n-layer and holes from the p-layer toward the opposite electrode, respectively.

B2.2.3 Photoconductive effect

When light strikes some kinds of semiconductor, electron–hole pairs are generated and their internal electric conductivity increases. This phenomenon is called the 'photoconductive effect'. Photoconductive detectors are divided into intrinsic detectors and extrinsic detectors doped with impurities.

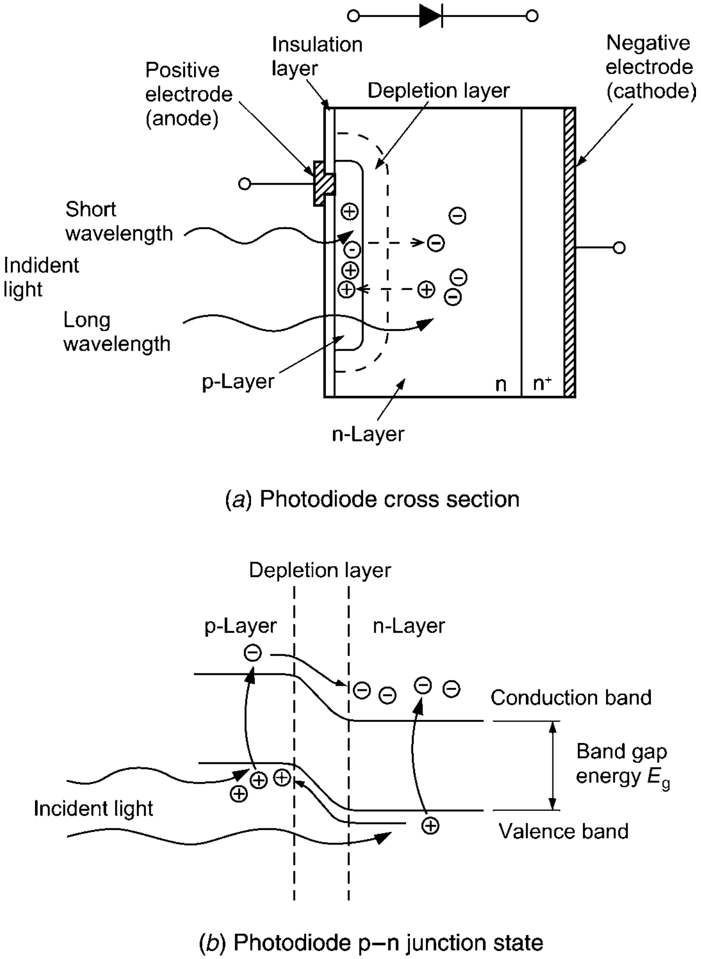

Insulation layer
Positive electrode (anode)
Depletion layer
Negative electrode (cathode)
Short wavelength
Indident light
Long wavelength
p-Layer
n
n+
n-Layer
(a) Photodiode cross section
Depletion layer
p-Layer
n-Layer
Conduction band
Band gap energy E_g
Incident light
Valence band
(b) Photodiode p–n junction state

Figure B2.3. (*a*) Photodiode cross section, (*b*) photodiode p–n junction state.

Figure B2.4 shows the operation models of photoconductive detectors. Figure B2.4(*a*) is a phenomenon called 'intrinsic photoconduction'. When photons $h\nu$ with energy greater than the energy band gap E_g in the forbidden band enter an intrinsic detector, electron–hole pairs (carriers) are generated and the number of conductive charges in the conduction band changes. Figure B2.4(*b*) shows the operation of an extrinsic detector doped with impurity atoms. In this extrinsic detector, an impurity level is formed at a relatively deep energy level in an n-type semiconductor. When the input photons have energy higher than the ionization energy E_1, they mainly act on the impurities and create free electrons and bound holes, and the free electrons and free holes contribute to changes in the electric conductivity. This phenomenon is called 'extrinsic photoconduction'. Since $E_g \gg E_1$, extrinsic detectors are used for IR detection at longer wavelengths when compared to intrinsic detectors.

Figure B2.5 shows how a photoconductive sensor is used to detect light. When light enters the photoconductive sensor, its internal resistance changes to produce an increase of electric current, ΔI, which is added to the constant bias current I_b. This change in the electric current is detected as a signal.

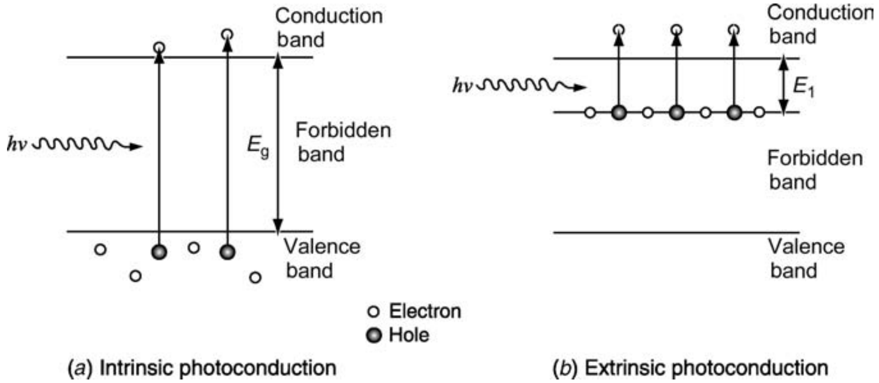

Conduction band
Forbidden band
Valence band
E_g
$h\nu$
Conduction band
E_1
$h\nu$
Forbidden band
Valence band
Electron
Hole
(a) Intrinsic photoconduction
(b) Extrinsic photoconduction

Figure B2.4. Operation models of photoconduction. (*a*) Intrinsic photoconduction, (*b*) impurity photoconduction.

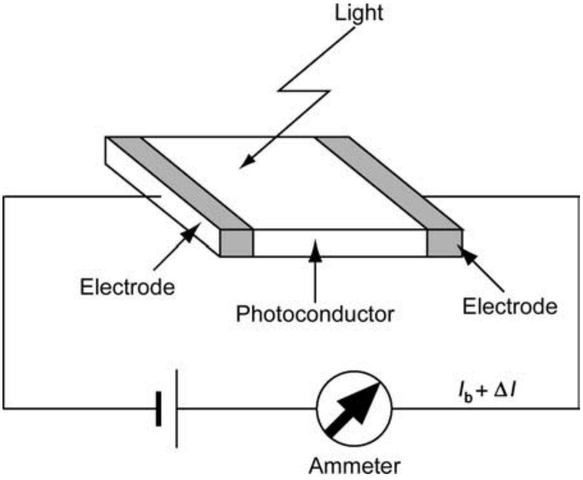

Light
Electrode
Photoconductor
Electrode
$I_b + \Delta I$
Ammeter

Figure B2.5. Photoconductive sensor operation.

Various types of material are used to fabricate photoconductive sensors, such as CdS for visible light detection and PbS, PbSe and InSb for IR detection. Although photoconductive sensors have the disadvantages of slow response speeds they are still widely used because of the useful features such as small size, light weight and wide spectral response range (from visible to far infrared range).

B2.3 Spectral response characteristics of photodetectors

It is essential to select an appropriate type of photodetector that matches the wavelength region of light to be measured. This section explains photodetectors used to detect wavelengths in the x-ray to IR region.

B2.3.1 Definition of photodetector sensitivity

A wavelength spectrum of light from x-rays to IR rays is shown in figure B2.6 along with the name of each spectral region and unit systems as a reference to the following descriptions in this chapter.

Here, the definitions of sensitivity used to evaluate photodetectors will be mentioned. If light at a certain intensity level (W) enters a photodetector and a certain amount of photocurrent (A) flows after photoelectric conversion, then the following terms are usually used to define the photoelectric sensitivity of the photodetector.

Radiant sensitivity or photosensitivity S

Radiant sensitivity or photosensitivity S is the photocurrent A divided by the incident light level W as represented in equation (B2.1).

$$S = \frac{A}{W}\,(\mathrm{A\,W^{-1}}). \tag{B2.1}$$

Quantum efficiency

Quantum efficiency (QE) is the number of electrons or holes extracted as photocurrent divided by the number of incident photons, and generally expressed in percent (%). QE and radiant sensitivity $S\,(\mathrm{A\,W^{-1}})$ have the relationship of equation (B2.2) at a given wavelength λ (nm).

$$\mathrm{QE} = \frac{S \times 1240}{\lambda} \times 100\,(\%). \tag{B2.2}$$

Noise equivalent power

This is the amount of light equivalent to the noise level of the photodetector. In other words, it is the light level required to obtain a signal-to-noise ratio (S/N) of 1. Noise equivalent power (NEP) is usually

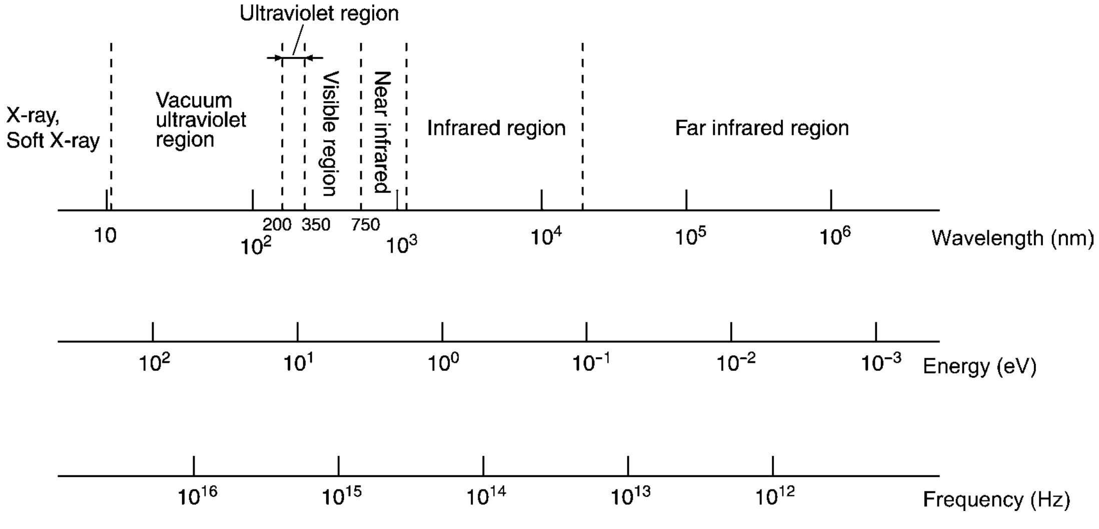

Ultraviolet region
X-ray,
Soft X-ray
Vacuum ultraviolet region
Visible region
Near infrared
Infrared region
Far infrared region
10
200
350
750
10^2
10^3
10^4
10^5
10^6
Wavelength (nm)
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
Energy (eV)
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
Frequency (Hz)

Figure B2.6. Wavelength spectrum of light and unit conversion scale.

defined as shown in equation (B2.3) using the wavelength λ_p of maximum sensitivity and a bandwidth of 1 Hz.

$$\mathrm{NEP} = \frac{\text{Noise current } (\mathrm{A\,Hz^{-1/2}})}{\text{Photosensitivity at } \lambda_p\,(\mathrm{A\,W^{-1}})}\ (\mathrm{W\,Hz^{-1/2}}). \tag{B2.3}$$

*D**

This indicates how much the S/N ratio of a detector is when radiant energy of 1 W enters the detector. *D** is normalized by a sensitive area of 1 cm^2 and noise bandwidth of 1 Hz so as to compare detector materials regardless of the size and shape of the detector element. An optical chopper is usually used to measure *D** for passing and interrupting a beam of incident light. *D** is normally expressed in a format of *D**(*A*, *B*, *C*) where *A* is the colour temperature (K) of the light source, *B* is the chopping frequency (Hz) and *C* is the noise bandwidth (Hz). *D** is therefore represented in units of $\mathrm{cm\,Hz^{1/2}}$, and the higher the *D** value, the better the detector. *D** is given by

$$D^* = \frac{(\mathrm{S/N})\Delta f^{1/2}}{PA^{1/2}} \tag{B2.4}$$

where *S* is the signal, *N* is the noise, *P* is the incident light energy ($\mathrm{W\,cm^{-2}}$), *A* is the sensitive area (cm^2) and Δf is the noise bandwidth (Hz). The following relation is established by *D** and NEP.

$$D^* = \frac{A^{1/2}}{\mathrm{NEP}}. \tag{B2.5}$$

To define photodetector sensitivity, radiant sensitivity and quantum efficiency are usually used for UV to visible photodetectors, while NEP and *D** are frequently used for IR detectors.

B2.3.2 X-ray to vacuum UV photodetectors

Light tends to behave as particles more actively in the x-ray to vacuum UV (VUV) region and absorption of light by substance also changes a great deal in some wavelength bands of this region. Because of this, it is important in this region to consider the stopping power (efficiency) versus light and the performance characteristics of window materials as well as photoelectric conversion sensitivity. Typical transmittance characteristics of window materials [5] used in the VUV, soft x-ray and x-ray regions are shown in figure B2.7. As can be seen from this figure, organic films and light metals such as Be (beryllium) and Al (aluminium) can be used as a window in the x-ray region at several kiloelectron volts or less, although there are currently no window materials available for some bands of the x-ray region. Figure B2.8 shows typical spectral transmittance of optical windows used in the UV to VUV region. Quartz glass and UV-transmitting glass are preferably used as transparent window materials in the UV region. MgF_2 (magnesium fluoride) and LiF (lithium fluoride) crystals are chiefly used as windows for detectors in the VUV region where significant absorption of light by oxygen and nitrogen occurs. Table B2.3 shows a list of typical photodetectors [6] used in the x-ray to VUV region.

Photodetectors utilizing gas ionization

Proportional counter tubes and Geiger–Mueller (GM) tubes as photodetectors using gas ionization have been around for quite a long time. Since the operating principles of these tubes are slightly different from each other, proportional counter tubes are capable of detecting both radiation energy and dose rate, while GM tubes are used only to measure dose rate. Due to the transmission bands of the window material (Be window) and fill-gas absorption characteristics, proportional counter tube applications are

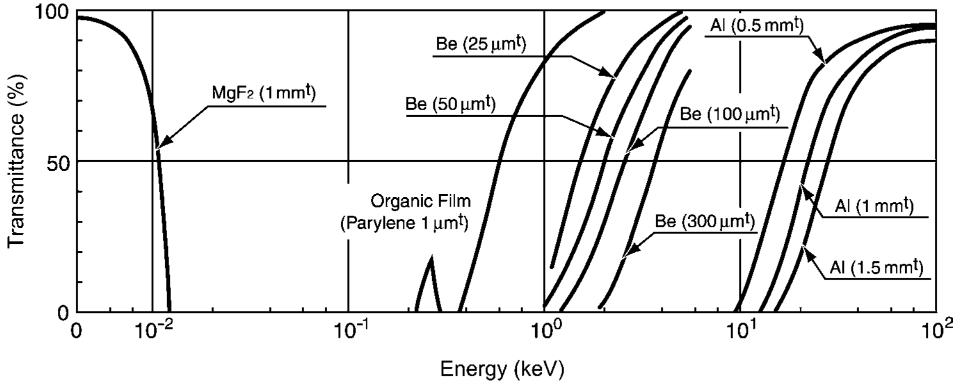


Figure B2.7. Transmittance of window materials used in VUV, soft x-ray and x-ray regions.

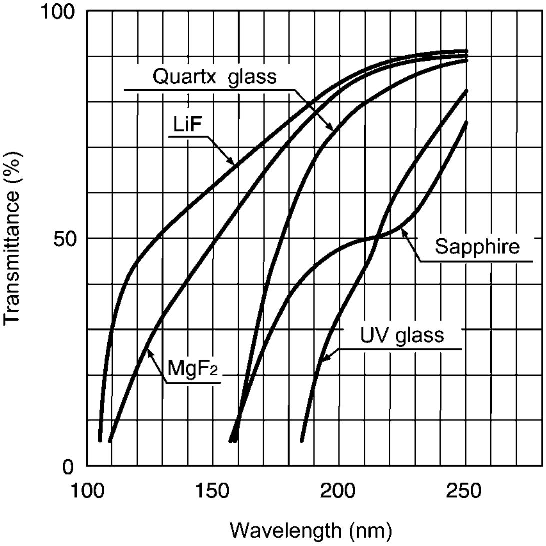


Figure B2.8. Spectral transmittance of optical windows used in VUV to UV region.

Table B2.3. Typical photodetectors used in x-ray to vacuum UV regions.

Detecting principle	Detectors	Detectable region	Energy (wavelength) region
Gas ionization	Proportional counter tube, GM tubes, MWPC	X-rays, γ-rays	A few keV to 1 MeV
Scintillation	Scintillator + PMT (or photodiode)	X-rays, γ-rays	A few keV to several dozen MeV
	X-ray image intensifiers	X-rays	20–150 MeV
Photoelectric effect	Semiconductor detectors	X-rays, γ-rays	A few keV to several dozen MeV
	PMT without window	Soft x-rays	A few to 100 nm
	PMT with window	VUV	100 nm or more

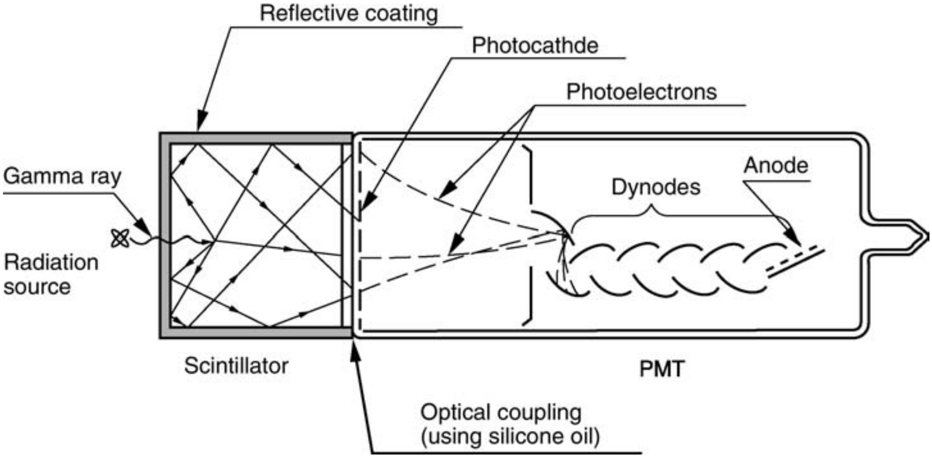


Figure B2.9. Scintillation photodetector for radiation measurement.

usually limited to a radiation energy region of 1–25 keV. Likewise, applications of GM tubes are determined by the transmission bands of the window material (Fe–Cr alloy) and by fill-gas absorption characteristics. GM tubes are therefore usable at a radiation energy of 50 keV–1 MeV. The multi-wire proportional counter (MWPC) is a position-sensitive proportional counter tube and is used as a two-dimensional detector in x-ray scattering, diffraction and synchrotron radiation experiments.

Scintillation photodetectors

In scintillation counting [7], a scintillator is used to convert radiation into visible light and this light is detected with a PMT or PD. Figure B2.9 shows a scintillation photodetector consisting of a scintillator coupled to a PMT. The amount of light emitted from a scintillator is proportional to the energy level of incident radiation, and this light emission is measured with a PMT or PD. Table B2.4 lists typical

Table B2.4. Typical characteristics and applications of scintillators.

	Density ($\mathrm{g\,cm^{-3}}$)	Relative emission intensity (NaI(Tl) = 100)	Emission time (ns)	Emission peak wavelength (nm)	Applications
NaI(Tl)	3.67	100	230	410	Survey meter, area monitor, gamma camera, SPECT
BGO	7.13	15	300	480	PET, x-ray CT
CsI(Tl)	4.51	45–50	1000	530	Survey meter, area monitor, x-ray CT
Pure CsI	4.51	<10	10	310	High energy physics experiment
BaF_2	4.88	20	0.9/630	220/325	TOF, PET
GSO:Ce	6.71	20	30	310/430	Area monitor, oil well logging
Plastic	1.03	25	2	400	Area monitor, neutron detector
LSO	7.35	70	40	420	PET
PWO	8.28	0.7	15	470	High energy physics experiment
YAP	5.55	40	30	380	Survey meter, compact gamma camera

characteristics and applications of major scintillators which are made of inorganic or organic materials. Scintillation counting is excellent in energy resolution when measuring radiation so it is used in nuclear medical equipment, nuclear physics experiments and oil well logging.

X-ray image intensifiers (IIs) [8] are a kind of scintillation photodetector. These are two-dimensional x-ray detectors having a phosphor-coated window for x-ray to visible light conversion and a photocathode sensitive to visible light. X-ray IIs can be fabricated with a large sensitive area, making them useful in x-ray medical equipment and industrial nondestructive inspection.

Photoconductive detectors

Among the photoconductive detectors, semiconductor radiation detectors [9] are classified by manufacturing method into the following groups:

(1) p–n junction type

(2) Surface barrier type

(3) Lithium (Si(Li)) drift type

(4) High-purity germanium type

Each type of these detectors makes use of the depletion layer that is formed when a reverse bias voltage is applied to the junction of p- and n-type semiconductors. Different photosensitive materials, size and window materials are used depending on the wavelength band (energy) to be detected.

Si(Li) drift type and high purity germanium type detectors are mainly used in the soft x-ray to x-ray region. Although the actual wavelengths that can be detected differ according to the detector material and structure, these can usually detect x-ray energy ranging from 1 keV to 10 MeV. These detectors are not easy to handle because liquid nitrogen cooling is required, but offer the advantage of high energy resolution. Recently, these detectors are available with a sensitive area from 2 or 3 up to 10 cm in diameter, and their detection sensitivity is constantly being improved. Multi-element detector arrays are also being developed.

Photocathodes (cathodes made of photoemissive materials) having external photoelectric effects are also used for detection in the UV to x-ray region. In a wavelength region where optically transparent windows are available, semi-transparent (transmission mode) photocathodes can be used, but at even shorter wavelength regions where proper window materials are unavailable, reflection mode photocathodes are exclusively used without windows. Typical photoemissive materials used to detect wavelengths shorter than the UV region are alkali halide metals, pure metals and metal oxide. Figure B2.10 shows typical spectral response characteristics of photoemissive materials that are usable in the soft x-ray region. These materials are applied to an x-ray to electron conversion surface such as the first stage of electron multipliers and also applied to a reflection mode photocathode by vacuum deposition onto the input edge surface of micro-channel plates (MCPs) (described later).

B2.3.3 UV detectors

Photodetectors used to measure radiant energy of UV rays can be divided into a photovoltaic type and photoemissive type.

As photovoltaic type semiconductor UV sensors, PDs using SiC, GaN, AlGaN and diamond thin films with a large band gap are now being developed [10]. In actual applications, UV sensors using a general-purpose PD combined with a UV filter are frequently used. Figure B2.11 shows a typical spectral response curve of a UV sensor, that can be obtained by the combination of a UV–visible sensitive

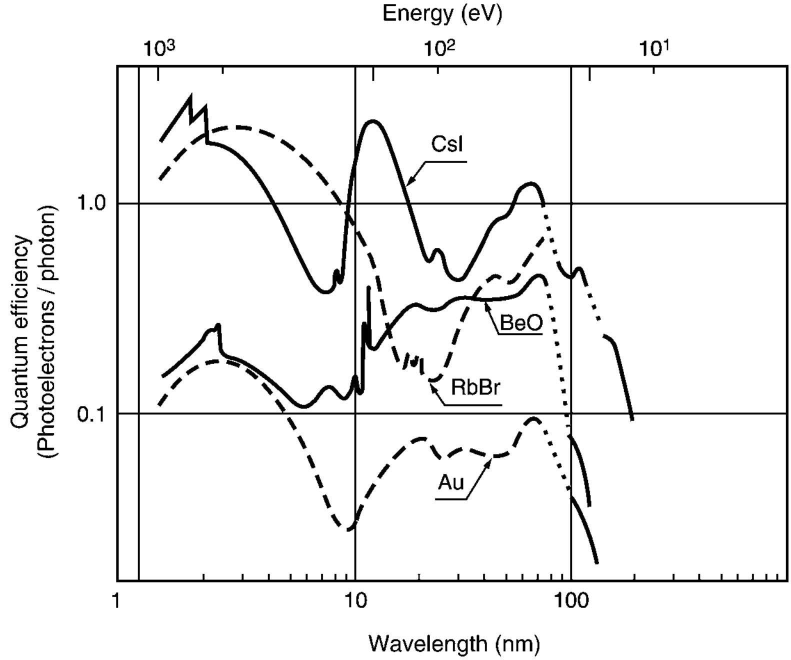


Figure B2.10. Typical spectral response characteristics of photoemissive materials in soft x-ray region.

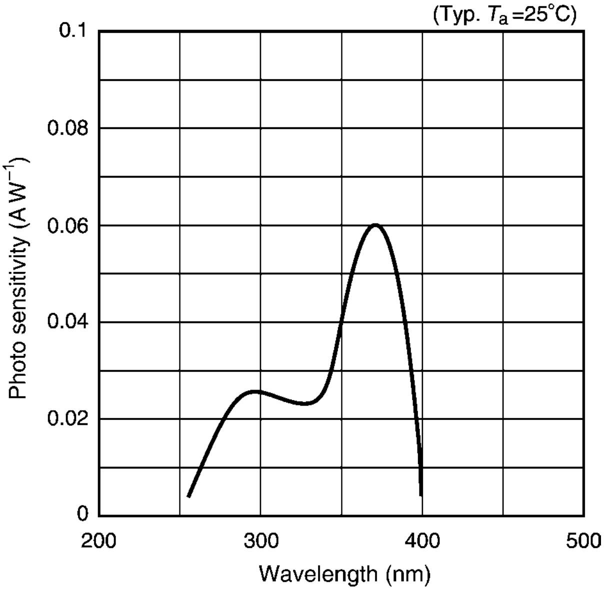


Figure B2.11. Typical spectral response curve of a semiconductor UV sensor.

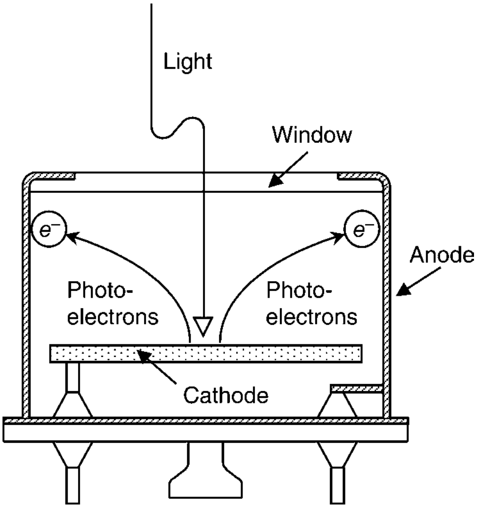


Figure B2.12. Construction of a phototube.

GaAsP PD and a UV-transmitting filter. Major applications of this type of UV sensor are UV measurement of sunlight, mercury lamp monitoring, etc.

Phototubes are also used in UV detection. They are made up of a photocathode using the external photoelectric effect and an anode for collecting photoelectrons emitted from the photocathode (see figure B2.12). Phototubes can be fabricated with various types of spectral response characteristic by changing the combination of photoemissive materials [11]. Figure B2.13 shows typical spectral response characteristics of photocathodes specifically selected and processed to have sensitivity only in the UV range. The spectral response characteristics on the short wavelength side are determined by absorption characteristics of window materials. Metals, alkali halide compounds and compound semiconductors having a large band gap are used as UV sensitive photocathode materials. Sensitivity in the longer wavelength region is determined by the band gap of each material. In addition to having various spectral response characteristics selectable by the combination of a window and photocathode materials, phototubes offer outstanding features such as high-speed response (nanosecond rise time) and low noise due to extremely high output impedance. Since UV phototubes are sensitive only in the UV region (insensitive to visible and longer wavelengths and often called 'solar blind'), they have found applications in colorimeters, pollution monitors, densitometers and UV laser detection.

B2.3.4 UV to visible photodetectors

Light is most actively utilized in this wavelength region. Compared to other regions of the electromagnetic spectrum, there are numerous types of photodetector available in the UV to visible region, in terms of both quantity and quality. Phototubes and PMTs utilizing the external photoelectric effect and PDs having the internal photovoltaic effect are most frequently used in this region.

Photodetectors utilizing external photoelectric effect (phototubes, PMTs)

Phototubes and PMTs are widely used in the UV to visible region. As stated earlier, these detectors use a semiconductor thin film that is a photoelectric converter surface called a 'photocathode'. Here, we will

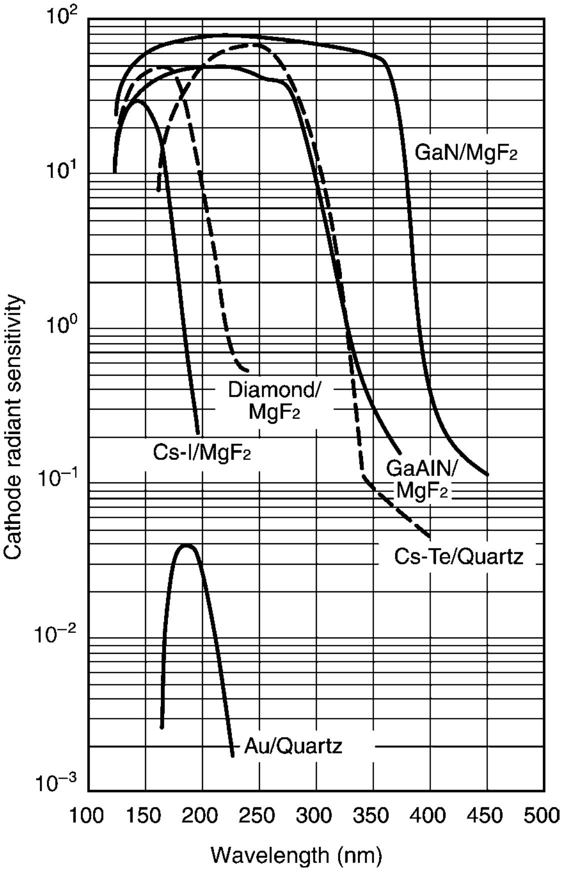


Figure B2.13. Typical spectral response characteristics of UV sensitive photocathodes.

briefly discuss the basic operating principle and operating method of a PMT. As shown in figure B2.11, a high voltage is applied between the photocathode and the anode. Photoelectrons emitted from the photocathode are accelerated and focused onto the first dynode (Dy1) to produce secondary electrons, which are then accelerated toward the subsequent dynodes (Dy2 to Dy*n*). When the accelerated electrons strike a dynode, secondary electrons are emitted with a secondary emission ratio (δ). Since this process is repeated to the last dynode (Dy*n*), the electrons are multiplied and increase. To make the operating circuit simpler, voltage dividing resistors are usually placed between the photocathode and the anode to distribute the supply voltage to each dynode, as shown in figure B2.14.

The secondary electron emission ratio δ is a function of the interstage voltage E between each dynode and is given by

$$\delta_1 = aE^k \qquad \text{(B2.6)}$$

where α is a constant, and k is a coefficient determined by the material and structure of an electrode (dynode) and has a value of 0.7–0.8.

The photoelectron current I_k (photocurrent that flows per unit light flux in lumens) emitted from the photocathode strikes the first dynode where a secondary electron current (I_{d1}) is emitted. At this point,

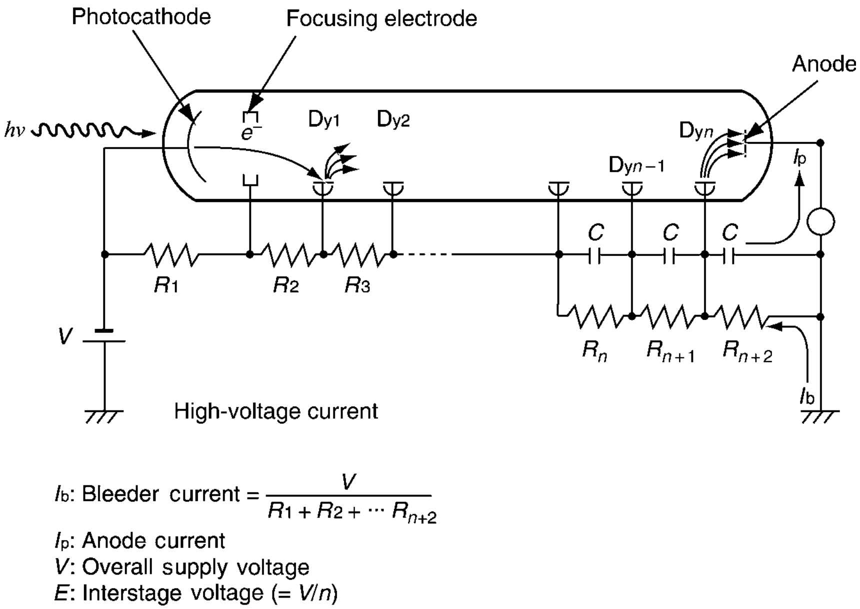

Photocathode
Focusing electrode
Anode
hν
e⁻
Dy1
Dy2
Dyn
Dyn−1
Ip
C
C
C
R1
R2
R3
Rn
Rn+1
Rn+2
V
Ib
High-voltage current
Ib: Bleeder current = V / (R1 + R2 + ··· Rn+2)
Ip: Anode current
V: Overall supply voltage
E: Interstage voltage (= V/n)

Figure B2.14. Photomultiplier tube operation circuit.

the secondary electron emission ratio δ of the first dynode is given by

$$\delta_1 = \frac{I_{\mathrm{d1}}}{I_{\mathrm{k}}}. \tag{B2.7}$$

These electrons are multiplied in a cascade process from the first dynode → second dynode → ··· → the nth dynode. The secondary emission ratio δ_n of the nth stage can be calculated by

$$\delta_n = \frac{I_{\mathrm{d}n}}{I_{\mathrm{d}(n-1)}}. \tag{B2.8}$$

The anode current I_{p} is given by

$$I_{\mathrm{p}} = I_{\mathrm{k}} \cdot \delta_1 \cdot \delta_2 \cdots \delta_n \tag{B2.9}$$

so that,

$$\frac{I_{\mathrm{p}}}{I_{\mathrm{k}}} = \delta_1 \cdot \delta_2 \cdots \delta_n. \tag{B2.10}$$

The product of $\delta_1, \delta_2, \ldots, \delta_n$ is called the gain (μ) and is given by

$$\mu = \delta_1 \cdot \delta_2 \cdots \delta_n. \tag{B2.11}$$

If the number of dynodes in a PMT is n, which is operated using an equal-division voltage divider, then changes in the gain μ versus the supply voltage V can be obtained by

$$\mu = \left(aE^k\right)^n = a^n \left(\frac{V}{n+1}\right)^{kn} = AV^{kn} \tag{B2.12}$$

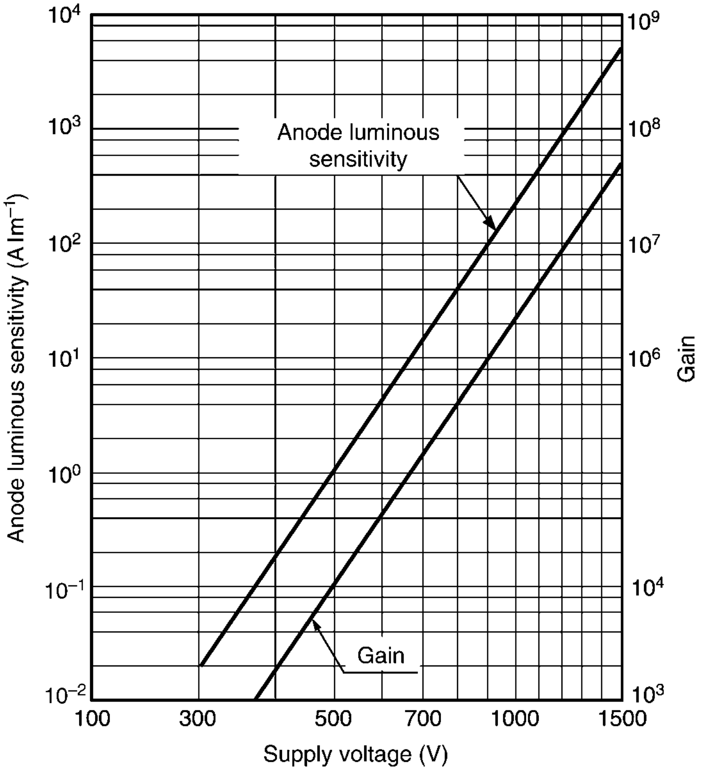


Figure B2.15. Luminous sensitivity and gain versus supply voltage.

where A should be equal to $a^n/(n+1)^{kn}$. From this equation, it is clear that the gain μ is proportional to kn exponential power of the supply voltage. Typical gain versus supply voltage is shown in figure B2.15 along with luminous sensitivity. Since this figure is expressed in a logarithmic scale for both horizontal and vertical axes, the slope of the straight line becomes kn. The gain increases with an increase in the supply voltage, so a high gain of up to 10^7 or more can be obtained in most cases. Figure B2.16 shows various types of dynode structure. The circular-cage dynode is mainly used in side-on PMTs and has a high gain and fast time response. The box-and-grid dynode has a high gain, the linear-focused dynode offers fast time response, and the Venetian blind dynode features a high gain and compactness. Recently, the mesh type dynode and MCP have been put to practical use as an electron multiplier of PMTs. These dynode types each have merits and demerits in terms of time response, dynamic range and gain, so they should be carefully selected according to the application. A PMT is usually operated with a supply voltage of 1000–2000 V in order to obtain a high gain of 10^6–10^7, and can be used in photon counting mode (explained later). PMTs have excellent capability for low-light-level detection and also deliver extremely fast response with rise times of nanoseconds (PMTs using normal dynode types) to subnanoseconds (PMTs using MCP).

PMT spectral response is determined by the photocathode material on the long wavelength side and the transmittance of the window material on the short wavelength side. Typical photocathode spectral response characteristics are shown in figure B2.17. Up until now, photocathode sensitivity has been limited to the UV to near IR region. However, recent developments in semiconductor crystal materials [12]

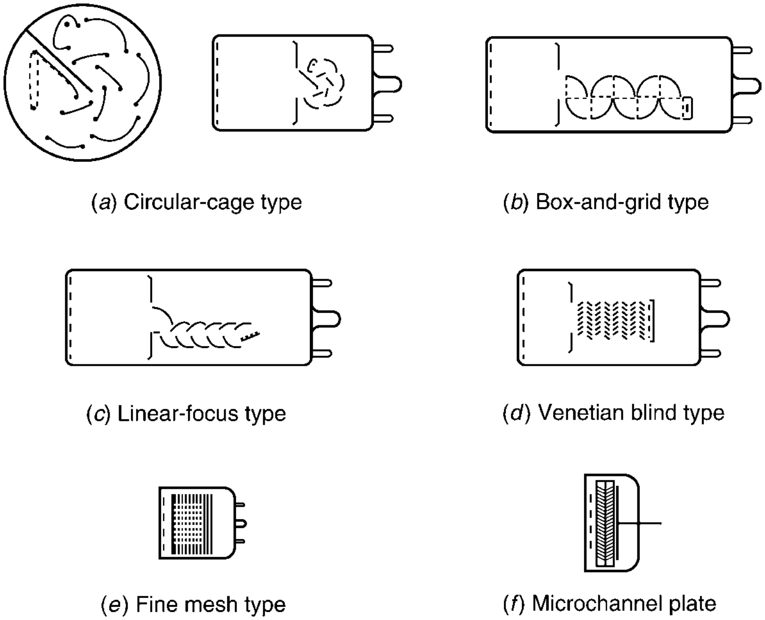

(a) Circular-cage type
(b) Box-and-grid type
(c) Linear-focus type
(d) Venetian blind type
(e) Fine mesh type
(f) Microchannel plate

Figure B2.16. Types of dynode. (*a*) Circular-cage type, (*b*) box-and-grid type, (*c*) linear focus type, (*d*) Venetian blind type, (*e*) fine mesh type, (*f*) microchannel plate.

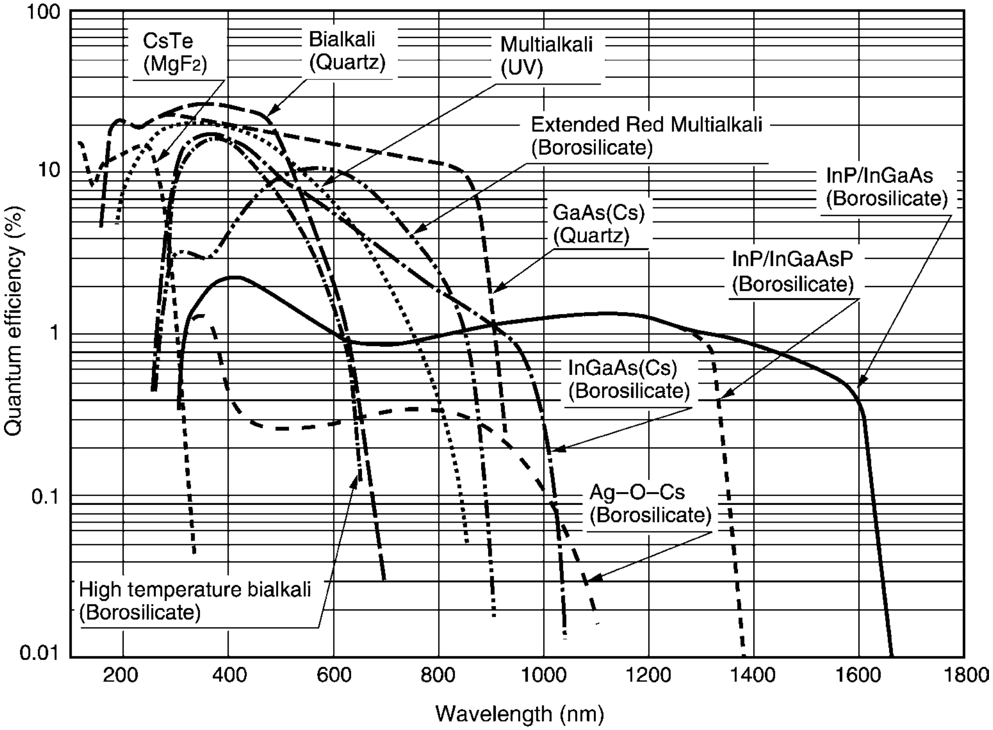


Figure B2.17. Typical photocathode spectral response characteristics.

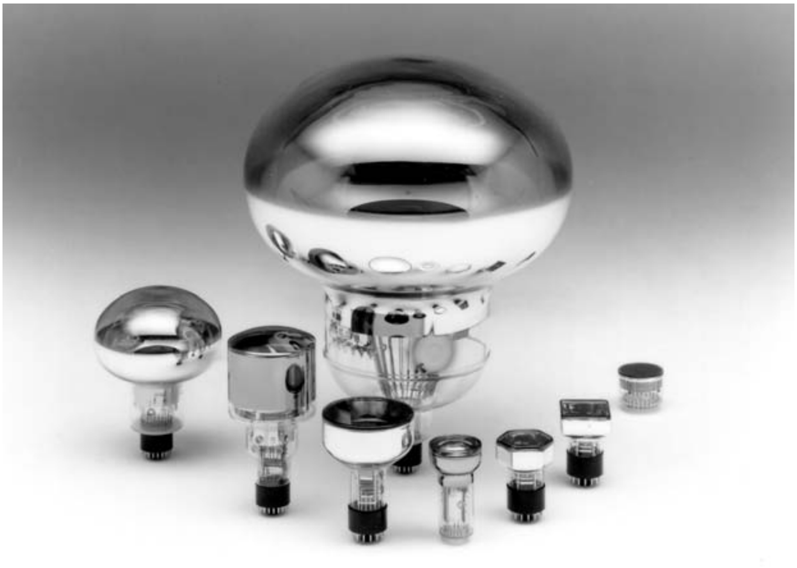


Figure B2.18. Typical photomultiplier tube products.

have extended photocathode sensitivity up to the IR region. Figure B2.18 shows a photograph of typical PMT products.

Photodetectors utilizing internal photovoltaic effect (PDs, etc)

Various types of semiconductor sensor have also been developed and used for light detection in the UV to visible region. Among these, PDs [4] are most extensively used. PDs can be classified according to their manufacturing method and structure as shown in figure B2.19. In the p–n type, light falling on a PD is absorbed in the depletion layer near the internal p–n junction after passing through the p^+-layer on the PD surface and is converted into an electrical signal of electron–hole pairs. The p–n type PDs have a planar structure that allows a relatively large active area with better sensitivity not only in the visible to IR region, but also in the UV region. However, these have a larger junction capacitance so response speed is limited. The pin type PDs have a high resistance i-layer formed between the p- and n-layers. This i-layer reduces the junction capacitance so response time is improved while maintaining high sensitivity. This type of PD exhibits an even faster response time when used with a reverse voltage applied and so is designed with low leak current.

Schottky type PDs have a structure in which a thin gold (Au) coating is sputtered onto the surface of an n-type semiconductor to form a Schottky effect p–n junction. Since the distance from the outer surface to the junction is small, UV sensitivity is high. Avalanche photodiodes (APDs) are used with a reverse voltage applied to the p–n junction so as to form a high electric field within the depletion layer. When light enters in this state, the generated electrons are accelerated by the electric field and collide with atoms to produce secondary electrons. This process occurs repeatedly so signals are amplified. This phenomenon is known as the avalanche effect and is ideal for detecting low level light.

Typical spectral response characteristics of PDs are shown in figure B2.20. Various types of PD are currently available with sensitivity in the UV, visible and also near IR regions. Recently, InGaAs PDs

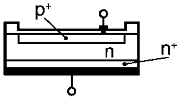
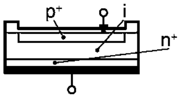
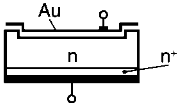
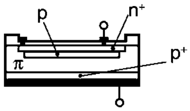
Type	Structure
p–n type	
pin type	
Schottky type	
Avalanche type (Reach-through type)	

Figure B2.19. Types of photodiode.

have become widely used as promising receivers in IR optical communications. Semiconductor photodetectors are small and compact yet have high sensitivity, so they are now used in large numbers in general electronics products and also other diverse applications. Figure B2.21 shows a photograph of typical PD products.

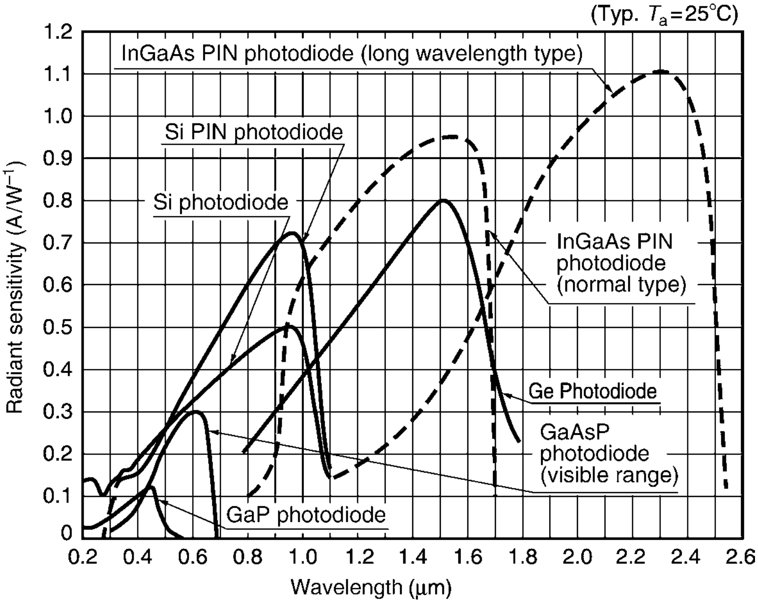


Figure B2.20. Typical spectral response characteristics of photodiodes.

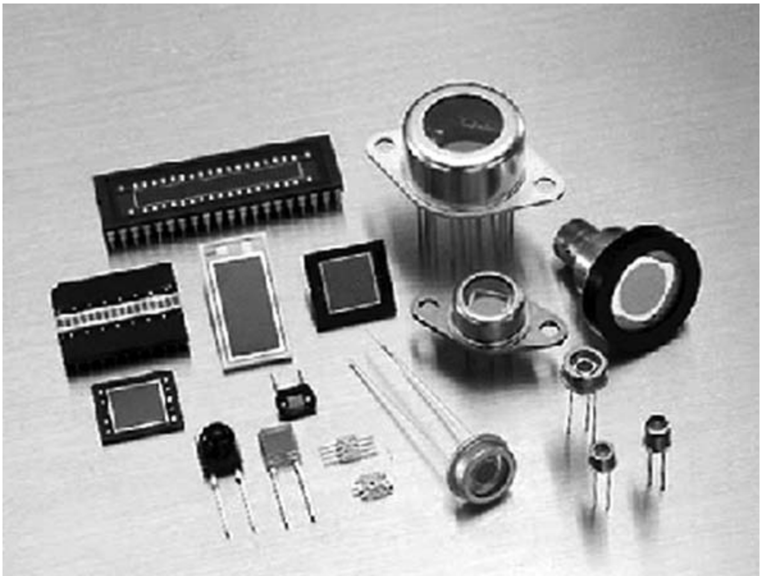


Figure B2.21. Typical photodiode products.

B2.3.5 IR detectors [13]

IR radiation covers the electromagnetic spectrum at wavelengths from 0.8 μm to 1 mm. The wavelength region of 0.8–3 μm is called the near IR, the wavelength region of 3–15 μm the middle IR and the wavelength region of 15 μm–1 mm the far IR.

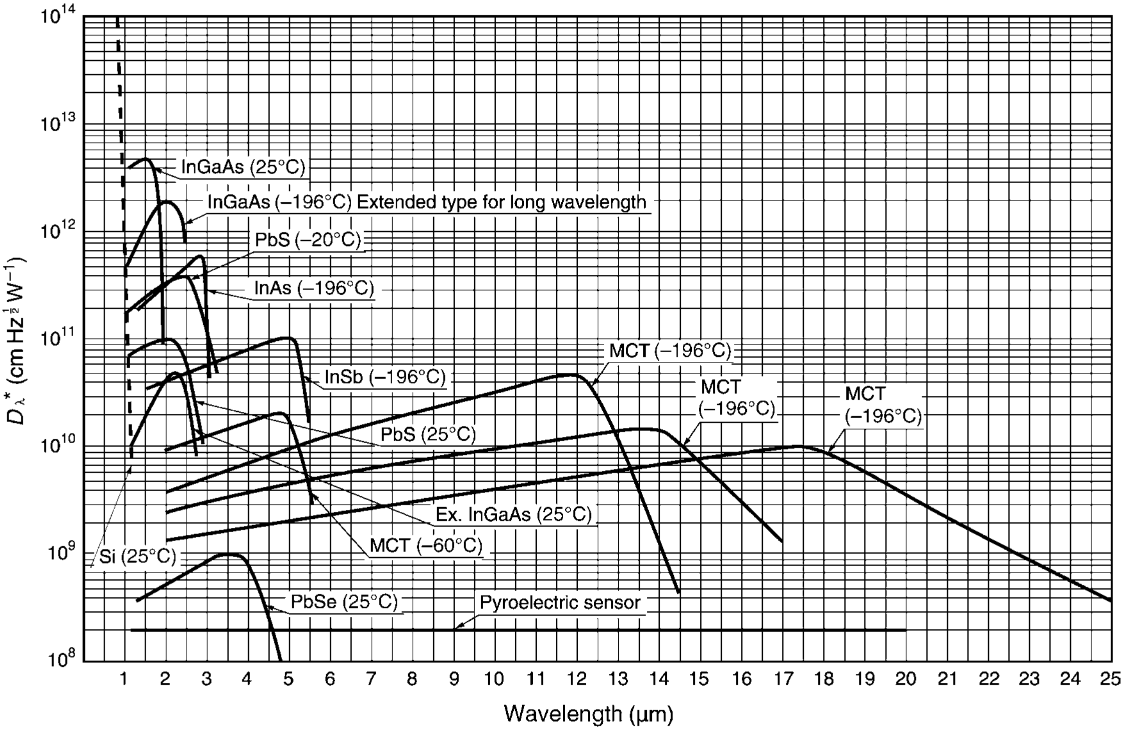
Table B2.5 shows types of IR detector and their characteristics [14]. In the IR region, thermal type and quantum type (photovoltaic and photoconductive) detectors are commonly used. Thermal type detectors make use of IR energy as heat, their responsivity is independent of wavelength and cooling is not required. However, their response speed is slow and detection sensitivity is not so high. Quantum type detectors, on the other hand, have higher responsivity and faster time response, while their responsivity depends on wavelength. Quantum type detectors often have to be cooled to ensure more stable operation.

Figure B2.22 shows typical spectral response characteristics of IR detectors. As can be seen from this figure, IR detectors using various kinds of photoelectric material have been developed to cover a broad spectral range from the near IR to far IR radiation. It should be noted that spectral responsivity of InGaAs, Ge and InSb detectors shifts to shorter wavelengths by cooling the detector element, while spectral responsivity of PbS, PbSe and MCT (HdCdTe) detectors shifts to longer wavelengths by cooling. Pyroelectric detectors are thermal type IR detectors made of pyroelectric materials such as $LiTaO_3$, TGS and PZT. Unlike quantum type detectors, pyroelectric detectors operate at room temperatures and their responsivity is not dependent on wavelength. One of their applications, recently the focus of much attention, is the 'human body sensor' that is finding wide applications in many fields such as home automation, security and energy saving.

IR detectors are used in diverse application fields including general electronics, security, disaster prevention, industrial measurement, communications, remote sensing, medical equipment and analytical instruments.

Table B2.5. Types of IR detector and their characteristics.

	Type of detector		Detectors	Spectral response range (μm)
Thermal type	Thermocouple, thermopile			
	Bolometer		Golay cell	Depends on window material
	Pneumatic cell		Capacitor microphone	
	Pyroelectric sensor		PZT, TGS, $LiTaO_3$	
Quantum type	Intrinsic type	Photoconductive type	PbS	1–3.6
			PbSe	1.5–5.8
			HgCdTe	2–16
		Photovoltaic type	InGaAs	0.7–1.7
			InAs	1–3.1
			InSb	1–5.5
			HgCdTe	2–16
	Extrinsic type		Ge:Au	1–10
			Ge:Hg	2–14
			Ge:Cu	2–30
			Ge:Zn	2–40
			Si:Ga	1–17
			Si:As	1–23



Hamamatsu Photonics; Compound Semiconductor Photosensors, Cat. No. KIRD0002E01(2001)

Figure B2.22. Typical spectral response characteristics of IR detectors.

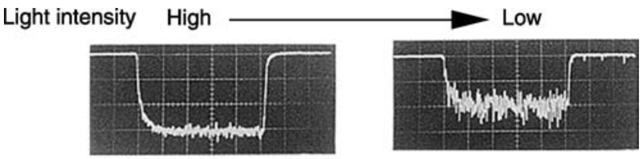
B2.4 Photodetector signal processing methods and S/N ratio

To operate photodetectors in actual applications, optimum detectors must first be selected by considering the light wavelength to be detected, the required time response and S/N ratio. In addition, the signal processing circuit connected to photodetectors must also be optimized. In particular, when the incident light intensity is very low, it is important to take countermeasures against external noise and design measurement systems that will maintain a satisfactory S/N ratio.

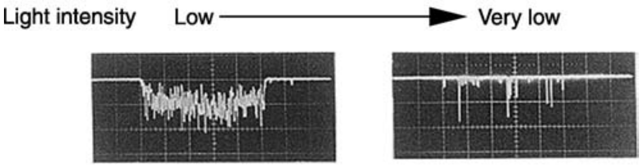
B2.4.1 Incident light intensity and signal processing method

Figure B2.23 shows output signal waveforms of a photodetector (a PMT is used here) observed on an oscilloscope while changing the intensity of light emitted from a pulse-driven LED. When the light intensity is high, the photoelectron pulses generated after photoelectric conversion overlap each other and create an analogue waveform as shown in figure B2.23(*a*). When the light intensity is reduced slightly, the output waveform will contain more AC components than DC components like those shown in figures B2.23(*a*)–(*c*). If the light intensity is reduced further, the output signal will be discrete pulses as shown in figure B2.23(*d*). This is the so-called 'photon counting region' (digital count mode).

In this way, the output signal waveform differs depending on the incident light intensity so the subsequent signal processing [15, 16] may use various methods. Typical optical measurement methods are shown in figure B2.24. The DC measurement method in (*a*) amplifies DC components from the photodetector and detects them through a low-pass filter. This method is used in optical measurement at rather high intensity and has been extensively used for many years. In the AC or pulse measurement method in (*b*), only the AC components in the output are extracted via a capacitor, amplified by a pulse amplifier, and converted into digital signals by a high-speed AD converter. This method is frequently used for demodulating pulsed light signals of wide bandwidth such as in optical communications. The photon counting method shown in (*c*) is a pulse counting method in which photoelectron pulses from the photodetector are amplified one by one, and only the pulses with an amplitude higher than the preset discrimination pulse height are counted as signals. Though not shown in figure B2.24, there are other optical measurement methods suitable for low-light-level detection even in applications subject to excessive noise. These include the boxcar method and lock-in detection method used in conjunction with an optical chopper in spectrophotometry.



(*a*) Analogue mode (*b*) Analogue mode



(*c*) Analogue mode (*d*) Digital mode

Figure B2.23. Signal waveforms observed on an oscilloscope when light intensity is changed. (*a*) Analogue mode, (*b*) analogue mode, (*c*) analogue mode, (*d*) digital mode.

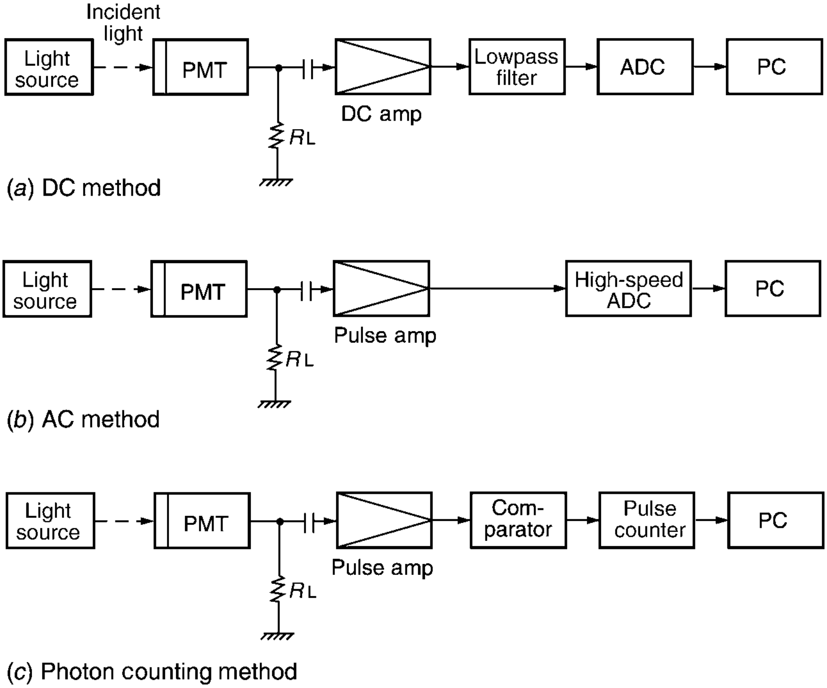

Incident light
Light source
PMT
DC amp
Lowpass filter
ADC
PC
RL
(a) DC method
Light source
PMT
Pulse amp
High-speed ADC
PC
RL
(b) AC method
Light source
PMT
Pulse amp
Com-parator
Pulse counter
PC
RL
(c) Photon counting method

Figure B2.24. Typical light measurement methods. (*a*) DC measurement, (*b*) pulse measurement, (*c*) photon counting.

B2.4.2 Photon counting method

Here, we will discuss specific circuit configurations used to perform photon counting [17, 18] which is an effective technique for light detection at extremely low light levels.

Figure B2.25 shows a typical circuit configuration for photon counting and a pulse waveform obtained at each circuit. In a PMT input photons are converted into photoelectrons which are then

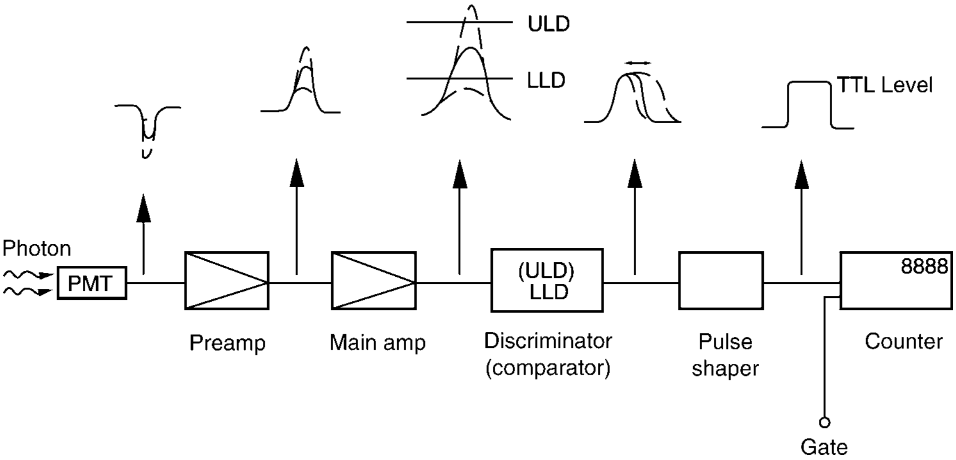

ULD
LLD
TTL Level
Photon
PMT
Preamp
Main amp
(ULD) LLD
Discriminator (comparator)
Pulse shaper
8888
Counter
Gate

Figure B2.25. Circuit configuration for photon counting.

multiplied by the dynodes up to 10^6–10^7. The multiplied pulses output from the PMT are converted into voltage pulses by a wide-band amplifier and amplified. These voltage pulses are fed to the discriminator and counted by a counter via a pulse shaper. The discriminator usually employs a comparator IC that compares the input voltage pulses with the preset reference voltage (threshold level) and eliminates those pulses with amplitudes lower than this value. In general, the lower level discrimination (LLD) is set at the lower pulse height side and the upper level discrimination (ULD) is set at the higher pulse height side to eliminate noise pulses with higher amplitudes. The counter usually has a gate circuit to set the desired measurement timing and intervals. This photon counting method is most effective in detecting extremely low level light.

To enable photon counting, photodetectors must have the following performance characteristics.

(1) Photodetectors must have adequate gain. Usually a gain of 10^5 or more is required in consideration of the relation between the next circuit's amplifier noise level and noise index.

(2) High quantum efficiency at wavelengths to be measured, and few noise pulses in order to attain a high S/N ratio.

(3) Narrow pulse height distribution.

(4) Large photosensitive area (when measuring diffused light).

PMTs satisfy almost all of the above prerequisites, so they are widely used in low-light-level photometric equipment such as photon counters. In recent years, PMTs for photon counting have been improved significantly, leading to development of PMTs [19] that deliver an exceptionally low noise pulse count around 0.1 cps. This allows measurement at extremely low light levels equivalent to nearly 1 photon per second.

In fluorescence lifetime measurement, changes in the light emission persisting only a very short time must also be measured along with the photon counting, so sophisticated techniques called time-resolved photon counting and time-correlated photon counting are used.

B2.4.3 Signal-to-noise ratio

In optical measurements, S/N ratio [20, 21] is a critical factor in determining the lower detection limit of photodetectors.

Figure B2.26 shows a typical model of a photodetector connected to an external circuit. Light is converted into electrons in the photoelectric section of the photodetector and the electrons are amplified if the photodetector has an electron multiplier function. The output from the photodetector is then amplified by the externally connected electronic circuit and extracted as an output signal.

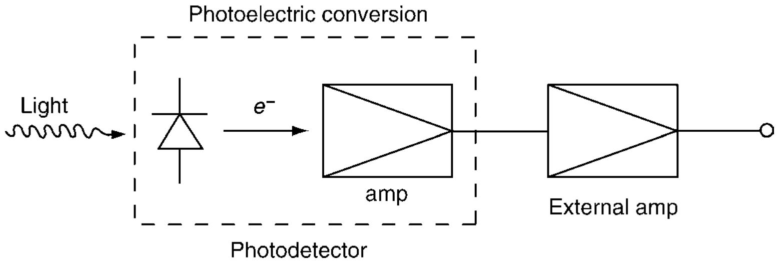


Figure B2.26. Typical model of photodetector connected to an external circuit.

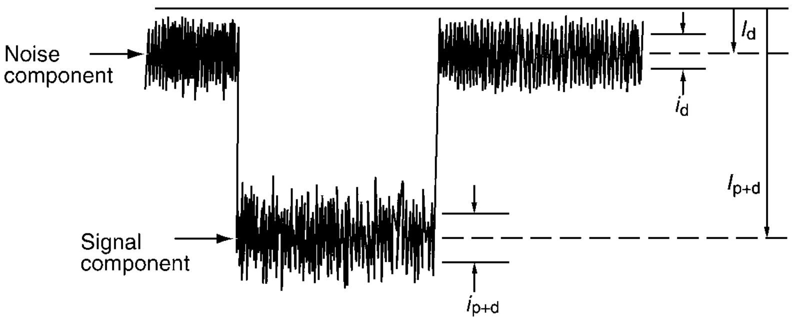


Figure B2.27. Analogue output waveform obtained from a photodetector.

Figure B2.27 shows an analogue output waveform generally obtained from a photodetector. This output contains both signal components produced by the incident light and noise components. These components can be defined as follows:

Mean value of noise component: I_d

AC component of noise: i_d (rms)

Mean value of signal (including noise component): I_{p+d}

AC component of signal (including noise component): i_{p+d} (rms)

Using these factors, the S/N ratio is given by

$$S/N = \frac{I_p}{i_{p+d}} \tag{B2.13}$$

where I_p is the mean value of the signal component obtained by subtracting I_p from I_{p+d}.

The extent of degradation in S/N ratio during the multiplication process is commonly expressed in terms of the noise figure (NF). The NF is defined by

$$NF = \frac{(S/N)^2_{in}}{(S/N)^2_{out}}. \tag{B2.14}$$

In general, the S/N ratio in analogue measurement regions can also be calculated by

$$S/N = \frac{I_k}{\sqrt{2eF_1(I_k + I_d) + \dfrac{4kTF_2}{R_L M^2}}} \frac{1}{\sqrt{B}} \tag{B2.15}$$

where I_k is the mean value of photocurrent generated by signal, e is the electron charge, F_1 is the noise figure of the photodetector, I_d is the mean value of dark current, k is Boltzmann's constant, T is the absolute temperature, F_2 is the noise figure of the externally connected amplifier, R_L is the equivalent resistance for connection to the externally connected amplifier, M is the gain of the photodetector, and B is the bandwidth of the entire system. The numerator in equation (B2.15) is the signal component and the denominator is the sum of the AC components (including the signal and noise) and the noise generated in the externally connected electronic circuit. The NF indicates the increase in noise ratio caused during the multiplication process in the photodetector. Since PMTs have high gain, the noise in the externally connected electron circuit can be ignored in most cases. The NF is about 1.3 for PMTs and 2–3 for APDs.

The second term of the noise of the numerator can be mostly ignored in PMTs since they have a high gain which can be regarded as infinite ($M = \infty$). On the other hand, this noise cannot be ignored in PDs because they have no gain ($M = 1$). If the bandwidth is widened, the S/N ratio will degrade.

In photon counting, the average value of the counts of individual photoelectron pulses is treated as a signal and fluctuations in the count values as noise. The S/N ratio in photon counting is given by

$$\mathrm{S/N} = \frac{N_\mathrm{s}\sqrt{T}}{\sqrt{N_\mathrm{s} + N_\mathrm{d}}} \tag{B2.16}$$

where N_S is the average value for the signal pulse count per second, N_d is the average value for the noise pulse count per second, and T is the counting time in seconds.

When comparing the S/N ratio of the photon counting and analogue methods (equations (B2.15) and (B2.16)), the extent of the NF makes it clear that the photon counting method is superior to the analogue method.

B2.5 High-speed photodetectors

There has been a shift in recent years away from conventional methods and towards using light in many ways. For example, light is extensively used to analyse substances and properties, make measurements and control various devices. A variety of different observation and measurement techniques have been developed for high-speed light measurement [22, 23] and put to practical use. To analyse even higher-speed phenomenon and achieve higher-speed information processing, light sources producing even shorter pulses and demodulators (photodetector and electrical circuit) with high-speed time response [24] are demanded. Here, high-speed photodetectors will be described after touching briefly on general methods for high-speed light measurement.

B2.5.1 High-speed light measurement techniques

Typical methods for measuring short light pulse signals from high-speed phenomena are shown in table B2.6. The analogue measurement sampling technique is a method for finding the waveform of a high-speed PD or PMT on an oscilloscope or other device. The response boundaries of this method are determined by the response of the high-speed photodetector, and the frequency bandwidth of the oscilloscope. Current technology has achieved a time resolution of around 10 ps. Among the real time measurement methods, the streak method [25] has the best time resolution (picosecond to subnanosecond range).

The auto-correlation method attains a time resolution of 10 fs and excellent high-speed characteristics yet cannot find the optical waveform itself due to its measurement principle.

The time-correlated photon counting (TCPC) method [26], along with having a comparatively fast time resolution, also has high sensitivity to low-level light, and a wide dynamic range (10^5–10^6) but also has the drawback of a long measurement time.

B2.5.2 High-speed photodetectors

High-speed photodetector types and time response characteristics are shown in table B2.7. Besides showing whether or not the detector has an amplification mechanism, this table also reveals large differences in photodetector response speeds. The biplanar phototube has a rise time of 60 ps but no amplification mechanism so detection is limited to strong pulsed light. Conventional PMTs have

Table B2.6. Typical methods for measuring short light pulses.

Measurement method	Detectors	Time resolution	Features	
Analogue measurement sampling method	Photodiodes Biplanar phototubes	Nanoseconds	Merits:	Easy to use, relatively fast response
	Photomultiplier tubes MCP-PMT	Subnanoseconds	Demerit:	Narrow dynamic range
Streak method	Streak camera	Femtoseconds	Merit:	Ultra-fast response
Auto-correlation method	SHG correlator	Femtoseconds	Merit:	Measurement of ultra-fast phenomenon
			Demerit:	Optical waveform cannot be measured directly
Time-resolved photon counting method	Photomultiplier tubes	Subnanoseconds	Merits:	Fast response, wide dynamic range
			Demerit:	Longer measurement time than analogue method

Table B2.7. High-speed photodetector types and time response characteristics.

Detectors	Amplification	Response speed
Biplanar phototubes	No	60 ps
Photomultiplier tubes	Yes	A few to several dozen nanoseconds
MCP-PMT	Yes	25 ps
Photodiodes	No	Subnanoseconds to microseconds
APDs	Yes	Subnanoseconds to microseconds
Streak tubes	Yes	0.2–20 ps

response in the nanosecond range but the PMT using MCPs (MCP-PMT) has a time resolution of 25 ps and a gain of 10^6.

Semiconductor photodetectors are widely used in measurement, control and analysis equipment as well as optical communications. Time response characteristics of semiconductor photodetectors are determined by carrier transit time in the depletion layer, the delay time set by the CR time constant, and in the case of APDs the time response is additionally determined by the avalanche rise time. Figure B2.28 shows the effective area versus frequency characteristics of pin PDs. As shown in the figure, the photosensitive area of the pin PDs is nearly proportional to the frequency response. A response up to 300 MHz is shown for an area size of 1.5 mm in diameter but even smaller photosensitive areas have a better response of several gigahertz.

High-speed photodetectors are particularly indispensable in the optical communications field due to their large data communication capacity and high speed. Current technology has achieved frequency characteristics up to 10 GHz using a long-wavelength band InGaAs PD, yet photodetectors with response characteristics of 50–90 GHz on up to the terahertz range will be in actual use in future.

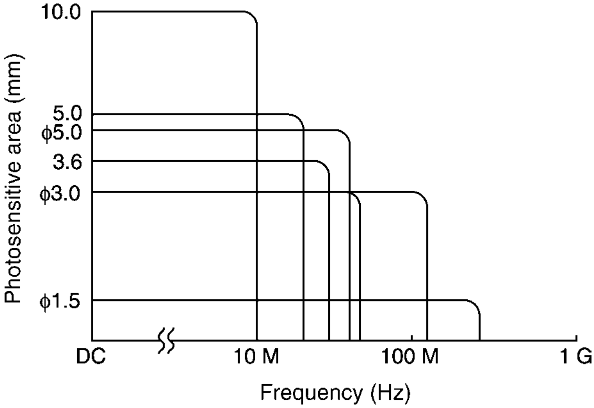


Figure B2.28. Effective area versus frequency characteristics of pin photodiodes.

High-speed devices called metal–semiconductor–metal photodetectors (or MSM-PD) using GaAs as the substrate are well known among semiconductor photodetectors. The MSM has a response of 20 ps (rise time) in actual operation and it was reported that a high-speed response in the subpicosecond [27] level could be attained.

Streak tubes have the highest speed among currently used photosensors, delivering a time resolution of 2 ps, but in recent years some have also attained the subpicosecond range. Development work on femtosecond streak tubes has attained a time resolution of 360 fs and just recently a time resolution of 180 fs has reportedly been reached [28]. Streak tubes are used for researching chemical reaction dynamics, measuring the semiconductor device relaxation process, and observation of super-high-speed phenomena such as implosions in laser nuclear fusion.

B2.5.3 Photodetectors for optical communication

Here we will discuss photodetectors used in optical communication. Digital communication [29] using light has become a useful tool in a wide range of applications mainly due to technical advances such as long-wavelength optical fibres (1.3 and 1.5 μm bands) having low transmission loss as well as rapid progress in wavelength multiplexing technology. High-speed digital communication using optical fibres requires photodetectors [30] with fast response speeds and high sensitivity that can be easily coupled to fibre optic cables.

Optical fibre communication

A typical block diagram for information transmission using an optical fibre is shown in figure B2.29. An information source is converted from electrical signals to optical signals (E/O conversion) after being

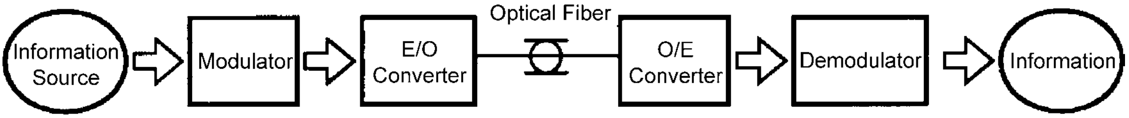


Figure B2.29. Typical block diagram for information transmission using an optical fibre.

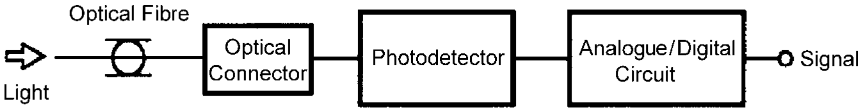


Figure B2.30. Typical configuration for fibre optic receiver.

modulated by various means and is then transmitted to the receiver side through the optical fibre. At the receiver, the optical signals are converted to electrical signals (O/E conversion) and then demodulated back into the original information. Optical fibre communication has the following advantages compared to conventional information transmission by electrical signals.

(1) High resistance to noise since there is no electromagnetic induction.
(2) Good insulation.
(3) Lower transmission loss means long-distance communication.
(4) Optical fibres are lightweight and take up less space.
(5) Huge bandwidth.

Performance required of photodetectors in optical fibre communication

Figure B2.30 shows a typical configuration for a fibre optic receiver. Light transmitted through an optical fibre is detected by a photodetector via an optical connector and converted to electrical signals. These signals are electrically processed by an analogue/digital circuit and then output. The following characteristics are required of photodetectors for use in optical fibre communication.

(1) High sensitivity at signal light wavelength.
(2) Time response fast enough for system transmission speed.
(3) Low noise level.
(4) High stability and long service life under continuous operation.

Besides optical connectors such as shown above for coupling between the optical fibre and photodetector, fibre-to-fibre direct coupling using pigtail type photodetector modules (see figure B2.31) is often used.

Receiver modules for optical fibre communication

In optical fibre communication, rather than as a discrete component, photodetectors generally come attached to optical fibres (pigtail type) or as the so-called optical receiver modules with preamplifiers built in for signal amplification. Various types of packaged receiver module are shown in the photograph [31] in figure B2.31. This photograph shows typical products including coaxial types with internal preamplifier, coaxial types without internal preamplifier, and a fibre jointed device (pigtail type). Long-distance communication uses laser diodes on the 1.3 and 1.5 μm bands, so InGaAs compound semiconductor photodetectors are used rather than silicon photodetectors. Due to the need for optical coupling with the fibres and high-speed response, APDs or PDs having a small sensitive area of 20–80 μm in diameter are generally used.

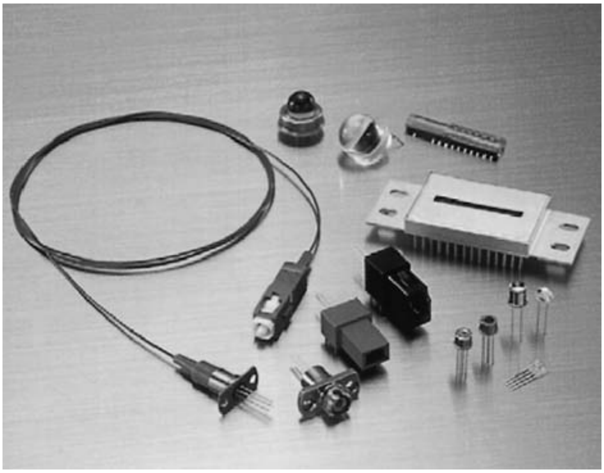


Figure B2.31. Various types of packaged receiver module.

Device characteristics of the most recent receiver modules of InGaAs PIN-PDs and InGaAs APDs are shown in table B2.8. The highest-speed modules are used in 10 Gbps long-distance communication.

Even further advances in optical communication technology are predicted. Technical innovations will expand communication now attainable at 10–40 Gbps and further future advances will take optical communication to speeds in the terabit per second range.

B2.6 Image sensors

Up to here in this chapter, we have defined photodetectors as sensors that only detect the intensity of light. We now however define image sensors as photodetectors that not only detect the intensity of light but also detect one-dimensional or two-dimensional information.

Table B2.8. Typical characteristics of receiver modules for optical communications.

	Active area (μm)	Sensitivity ($A W^{-1}$) Typical	Wavelength (nm)	Bandwidth fc (GHz) Minimum	Preamplifier	Package	Data rate
InGaAs PIN-PD	40	1.00	1550	8.0	Built in	17-pin mini butterfly	10 Gbps
	50	0.89 0.94	1310 1550	2.5	No	Coaxial	2.5 Gbps
	80	0.89 0.94	1310 1550	2.5	No	Coaxial	≤622 Mbps
InGaAs APD	20	0.7	1550	7.0	Built in	17-pin mini butterfly	10 Gbps
	50	0.94 0.96	1310 1550	2.5	Built in	Coaxial	2.5 Gbps
	50	0.94 0.96	1310 1550	1.0	No	Coaxial	≤622 Mbps
	50	0.94 0.96	1310 1550	2.5	No	Coaxial	2.5 Gbps

Table B2.9. Categories of image sensors.

Detecting principle	Readout method	Typical product	Sensitivity	Wavelength range
Image sensors				
External photoelectric (electron tube sensor)	Photoemissive	Image intensifier	Super-high sensitivity	VUV to near IR
	Photoconductive	Vidicon	Sensitive to invisible light	Visible to IR
Internal photoelectric (solid-state sensor)	Address type	MOS CID	High sensitivity	UV to near IR
	Charge transfer	CCD	High sensitivity	UV to near IR
Hybrid sensor		ICCD EBCCD	Super-high sensitivity	VUV to near IR

Table B2.9 shows categories of image sensors. Image sensors can be roughly classified into electron tube image sensors, solid state image sensors and hybrid image sensors combining both types. Electron tube image sensors are further divided into those using the external photoelectric effect or the internal photoelectric effect. IIs are typical electron tube image sensors using the external photoelectric effect. IIs are extremely high-sensitivity image sensors since they have an internal electron multiplier called MCP that delivers a multiplication factor (gain) of 10^3–10^6 times. IIs were primarily developed for low-light imaging such as at night time so they are still referred to as night vision tubes. In addition to low-light imaging applications, IIs are applied to high-speed shutter cameras because an electronic shutter that works in subnanoseconds can be implemented. Image pickup tubes or camera tubes are the best known among electron tube image sensors using the internal photoelectric effect. Among these, camera tubes called 'vidicon' have been used for many years in various applications from the UV, visible and IR regions by changing their photoconductive target materials (Sb_2S_3, PbO, CdSe).

Typical solid state image sensors are charge transfer type CCD image sensors and CID image sensors, and addressed type MOS image sensors. More recently, hybrid image sensors using an electron tube image sensor and a semiconductor image sensor have been developed and put to practical use.

Image sensors can also be classified by application into quantitative measurement such as for spectrophotometry and pattern recognition; and general imaging such as home video cameras. Since only limited space is available, this section will explain about image sensors designed for measurement applications.

B2.6.1 Electron tube image sensors

Camera tubes, the best known electron tube image sensors, have been used in many imaging applications. In recent years, however, camera tubes have been almost completely replaced by semiconductor image sensors and are now seldom used in the visible light region.

IIs [32, 33] are electron tube image sensors with ultra-high sensitivity and a high-speed electronic shutter mechanism.

Figure B2.32 illustrates the structure of a proximity-focused II. In operation of this type of II, an optical image from a low-light-level object is converted into photoelectrons by the photocathode, multiplied by the MCP and reconverted into an amplified optical image on the phosphor screen.

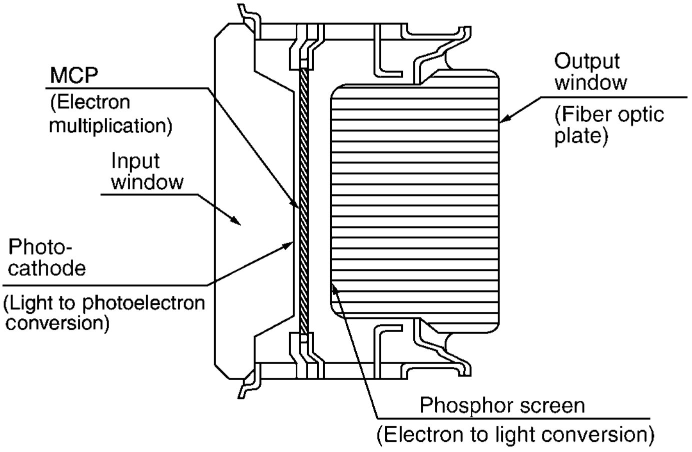
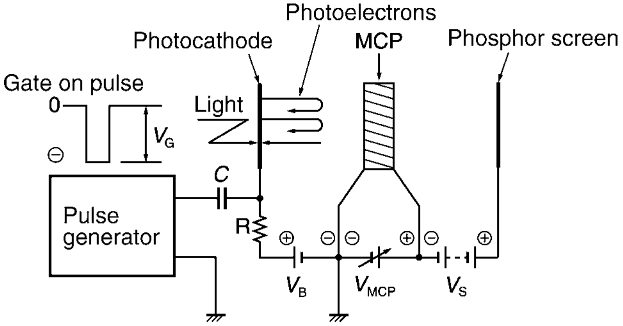

MCP
(Electron multiplication)
Input window
Photo-cathode
(Light to photoelectron conversion)
Output window
(Fiber optic plate)
Phosphor screen
(Electron to light conversion)

Figure B2.32. Structure of a proximity-focused image intensifier.

An MCP consists of a multitude of glass capillaries (channels) of 10–20 μm in inner diameter, fused together and formed into the shape of a thin disc of 0.1–1.0 mm thick. The inner wall of each channel is coated with a secondary emissive material with a proper resistance value, so each channel serves as an independent, nondiscrete secondary electron multiplier. For example, in an MCP of 18 mm in outside diameter, about 10^6 channels are arrayed in two dimensions so each channel corresponds to a pixel when used as a two-dimensional sensor. This MCP multiplication function allows IIs to have a high gain of 10^3–10^6 (depending on the number of MCP stages), so that they are used in night vision devices and low-level-light image sensors. Figure B2.33 is a circuit for driving a high-speed electronic shutter (gate) built into an II. By applying a negative pulse between the photocathode and the MCP, the photoelectrons emitted from the photocathode can be controlled so as to have an electronic shutter function [34] that turns the II operation on or off. In figure B2.33, the shutter turns on when a negative voltage pulse is applied to the photocathode. This electronic shutter is used at speeds of nanoseconds to microseconds in most applications, although special shutters are designed to operate at ultra-high speeds of 50 ps [35]. When viewing highly repetitive phenomena, phosphor materials [36] having a short decay time τ


Photoelectrons
Photocathode
MCP
Phosphor screen
Gate on pulse
0
⊖
V_G
Light
C
Pulse generator
R
⊕
⊖
⊖
⊕
⊖
⊕
V_B
V_{MCP}
V_S

Example: $V_B = 30V$, $V_G = 230$ V

Figure B2.33. Electronic shutter operation.

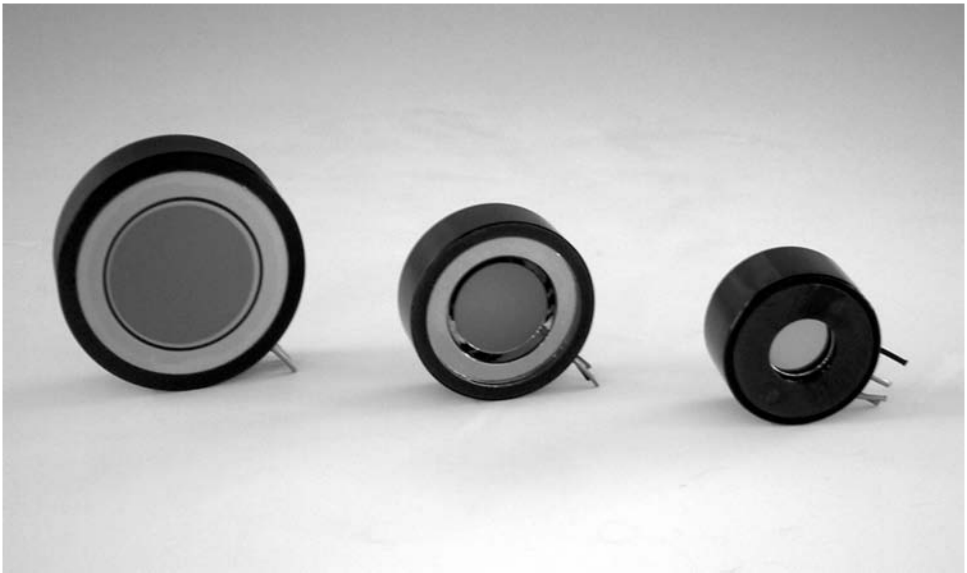


Figure B2.34. Image intensifier product examples.

designated P47 ($\tau = 0.1\ \mu s$), P46 ($\tau = 0.2\ \mu s$) or P24 ($\tau = 8\ \mu s$) are used. But, in general applications, P43 ($\tau = 1$ ms) is used because of high emission efficiency.

By selecting the window and photocathode materials, IIs can be used for imaging in various wavelength regions of the spectrum from soft x-rays to IR radiation. II products are shown in the photograph in figure B2.34.

B2.6.2 Semiconductor image sensors

Charge coupled device (CCD) image sensors and MOS image sensors are typical semiconductor image sensors [37, 38]. They are broadly divided by structure into one-dimensional arrays (linear image sensors) and two-dimensional array sensors (area image sensors). Semiconductor image sensors are generally mass-produced for image capturing applications but here we will discuss their use in making measurements.

Semiconductor image sensors are also grouped into address types and charge transfer types according to their scan method.

The address type shown in figure B2.35(*a*) is a two-dimensional array of pixels consisting of photoelectric elements such as PDs. Each pixel is respectively connected to a vertical and horizontal shift register. To read out signal charges from pixels, an external pulse is sent in sequence to the pixels, and the signal charge accumulated at the intersecting pixel is read out. In the charge transfer type shown in figure B2.35(*b*), signal charges of pixels comprised of PDs are sequentially transferred in parallel vertically, and also transferred one line at a time horizontally, and the signal charge then read out. MOS (metal-oxide semiconductor) image sensors are typical address type sensors, while CCD image sensors are typical charge transfer type sensors.

The CCD has high sensitivity, low noise and is capable of handling large numbers of pixels. The CCD is the most dependable yet least expensive device among currently available image sensors and is used in large numbers. MOS image sensors offer a large photosensitive area capable of handling a large

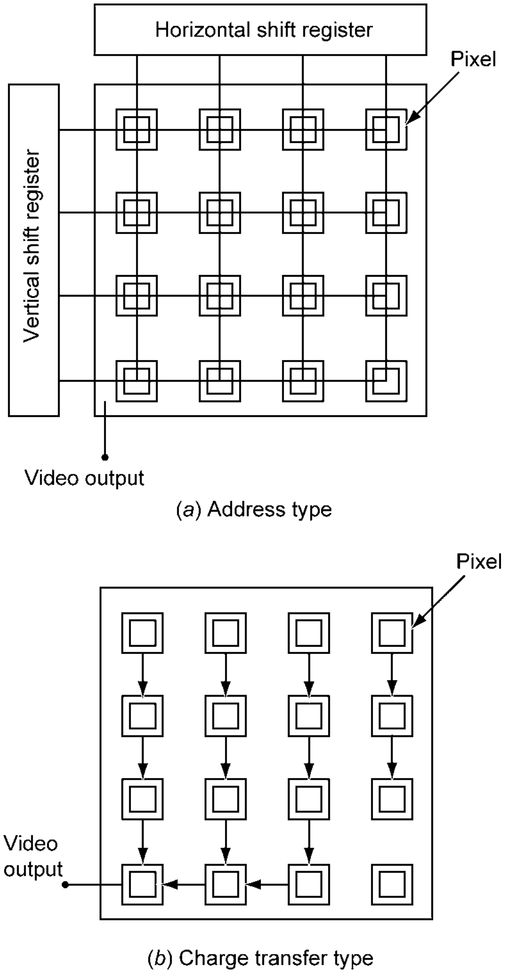

Horizontal shift register
Pixel
Vertical shift register
Video output
(a) Address type
Pixel
Video output
(b) Charge transfer type

Figure B2.35. Operating principle of semiconductor image sensors. (*a*) Address type, (*b*) charge transfer type.

storage charge and have low power consumption. One disadvantage of MOS image sensors is large capacitive noise compared to CCD, and the S/N ratio gets worse under low level light.

In this section, we cover image sensors intended mainly for optical measurement, so refer to other texts if seeking information on general-purpose image sensors such as for video cameras [39].

Linear image sensors [40]

Linear image sensors are image sensors segmented in only one dimension. A typical application of CCD linear image sensors is for facsimile readout and image sensors are also used in multichannel wavelength detectors in polychromators for spectrophotometry.

Figure B2.36 shows the structure of an n-MOS linear image sensor. The photosensitive area is a p–n junction PD consisting of a p-type silicon substrate on which an n-type diffusion layer is formed. The sensor has a photoelectric function to convert the optical signal into an electrical signal and a function to temporarily store the acquired signal charge.

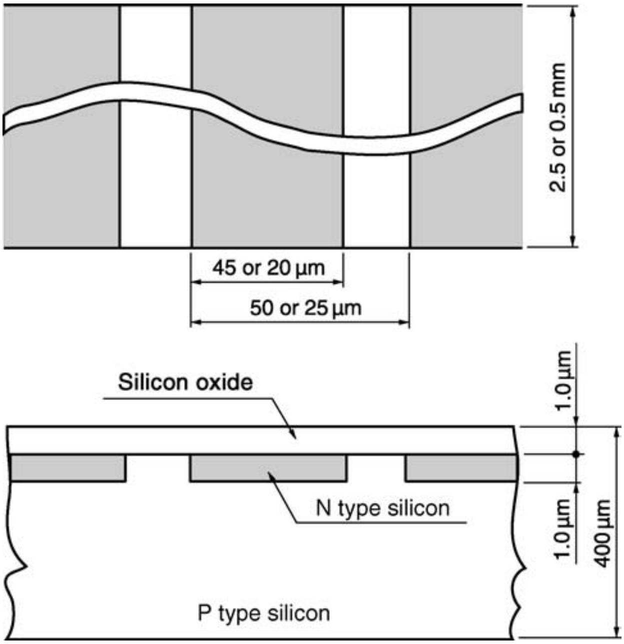


Figure B2.36. Structure of an n-MOS linear image sensor.

Table B2.10 shows typical structure and device characteristics of linear image sensors having a photoelectric surface made of InGaAs or silicon material. The photosensitive layers of InGaAs or silicon, respectively, match the visible-to-near-IR and the IR regions. Pixel size is several millimetres in height and up to 100 µm in width, with some hundreds to thousands of pixels fabricated as channels. Figure B2.37 shows an equivalent circuit of an n-MOS linear image sensor. The photosensitive area of the PD, as shown in the figure, is comprised of a switching section to read out the PD signal, and a shift register to address that switch. The linear image sensor using silicon may be fabricated to have a maximum of 2048 channels. These image sensors are usually used singly but may often be used in combination with an II described later on.

The image sensor using InGaAs is utilized for spectrophotometry in the IR on a wavelength of approximately 1–2.6 µm. As a photodetector for near-IR multichannel spectrophotometry this has

Table B2.10. Typical structure and device characteristics of linear image sensors.

Detector material	Window material	Spectral response range	Cooling	Pixel size	Number of pixels
Si	Quartz	200–1000 nm	Yes	50 µm × 2.5 mm 25 µm × 2.5 mm 50 µm × 0.5 mm 25 µm × 0.5 mm	256–2048
InGaAs	Sapphire	0.9–1.7 µm	No	50 µm × 250 µm	128–512
		0.9–1.7 µm	Yes	25 µm × 250 µm	
		1.2–2.6 µm	Yes	50 µm × 0.5 mm 25 µm × 0.5 mm	

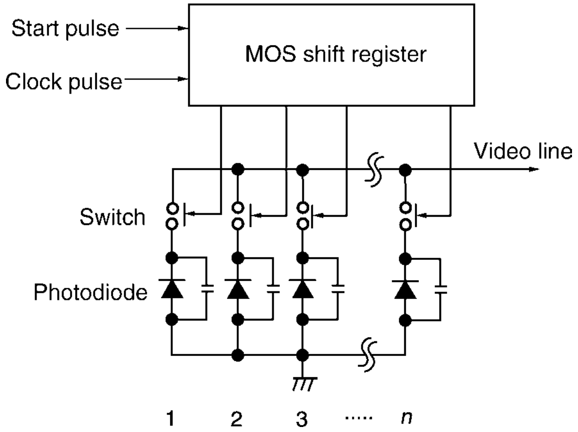


Figure B2.37. Equivalent circuit of an n-MOS linear image sensor.

applications in inspection of farm products by the IR absorption technique, radiation thermometers and nondestructive inspections, etc.

Area image sensors

Two-dimensional image sensors (area image sensors) for measurement applications often use CCD image sensors [41]. The transfer method, structure and characteristics of CCD image sensors are described below.

CCD image sensor transfer method. CCD image sensors can typically be grouped into one of the following three types of transfer method:

1. Frame transfer type: The frame transfer (FT) type CCD shown in figure B2.38 is comprised of two vertical shift registers made up of a photosensitive area and charge storage area, one horizontal shift register and an output section. A transparent electrode such as polycrystalline silicon is generally used as the metal electrode for the photosensitive area. The areas other than the photosensitive area are covered with a nontransparent electrode such as aluminium so that light will not enter those areas.

2. Full frame transfer type: The full frame transfer (FFT) type CCD in figure B2.39 is basically the FT type CCD without the charge storage area. Since there is no charge storage section, it must normally be used with some kind of external shutter mechanism. This limitation makes it difficult to use as a video camera so its use is mainly limited to measurement applications.

3. Interline transfer type: In the interline transfer (IT) type CCD shown in figure B2.40, the photosensitive area is formed separately from the transfer section. Here the vertical shift register is formed on the sides so as to enclose the PD that constitutes the photosensitive area. Because of this structure, the numerical aperture is low, so this type is not suited for measurement applications.

To sum up the above described methods, the IT type is used in video cameras while the FT and FFT types are mainly used in measurement applications.

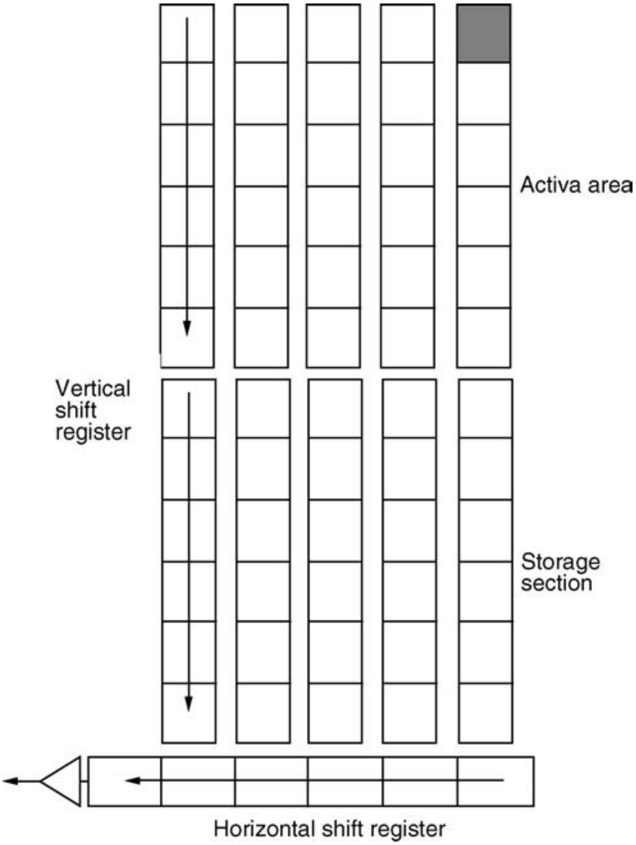

Activa area
Vertical shift register
Storage section
Horizontal shift register

Figure B2.38. Structure of a frame transfer type CCD.

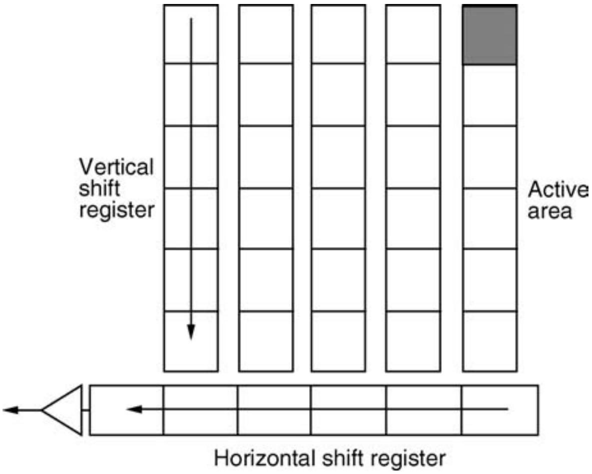

Vertical shift register
Active area
Horizontal shift register

Figure B2.39. Stucture of a full frame transfer type CCD.

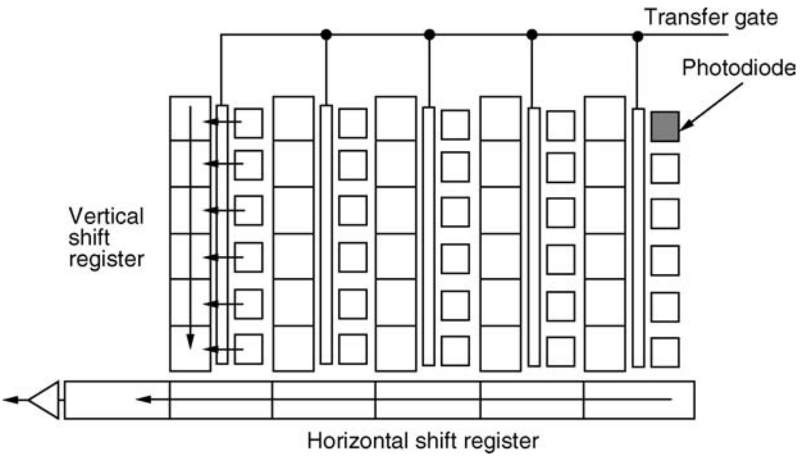


Figure B2.40. Structure of an interline transfer type CCD.

Structure of area image sensor for measurement work. To function effectively in measurement applications, the area image sensor must have high sensitivity, a wide dynamic range and be multichannel. Table B2.11 shows pixel formats of typical area image sensors used in measurement applications. The structure of an FFT (or FT) type CCD image sensor is generally used. A rectangular format is generally used for spectrophotometry as a precondition for binning (described later), and a square format is used for imaging. Actual CCD image sensor products are shown in the photograph in figure B2.41.

To improve measurement performance of the CCD image sensor, low noise is achieved by cooling (to −20°C), sensitivity is improved by back-illumination, and scanning is used in the binning operation.

Figure B2.42 shows the structure of a back-illumination CCD [42] (also called 'back-thinned CCD' since the backside of the CCD chip is thinned).

As shown in the figure, light enters from the front surface on a normal CCD. However, light loss occurs when light is input from the front due to absorption of light by the oxide film and the electrodes. In contrast, in the back-illumination CCD there is no such light loss, high sensitivity is obtained and a quantum efficiency as high as 90% is achieved. Even in the UV region, a high quantum efficiency of 40% or more is achieved. Spectral response characteristics of the back-illumination CCD are shown in figure B2.43, along with characteristics of the front-illumination CCD for comparison. A great increase in sensitivity can be seen in the UV region as well as in the visible region.

Table B2.11. Pixel formats of typical area image sensors used in measurement applications.

Applications	Number of pixels (H × V)	Pixel size
Spectrophotometry	532 × 64	24 μm × 24 μm
	1044 × 64	24 μm × 24 μm
Imaging	512 × 512	24 μm × 24 μm
	1024 × 1024	24 μm × 24 μm
	2044 × 2033	9.0 μm × 9.0 μm
	3072 × 2048	9.0 μm × 9.0 μm

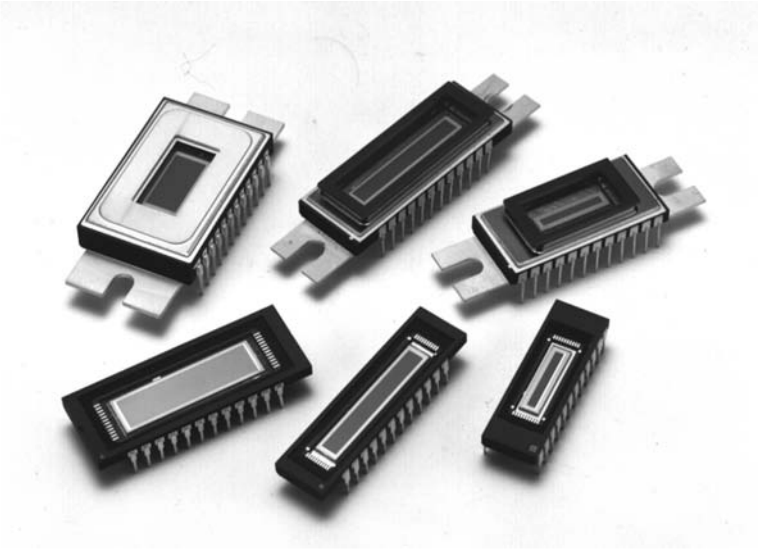


Figure B2.41. CCD image sensor products.

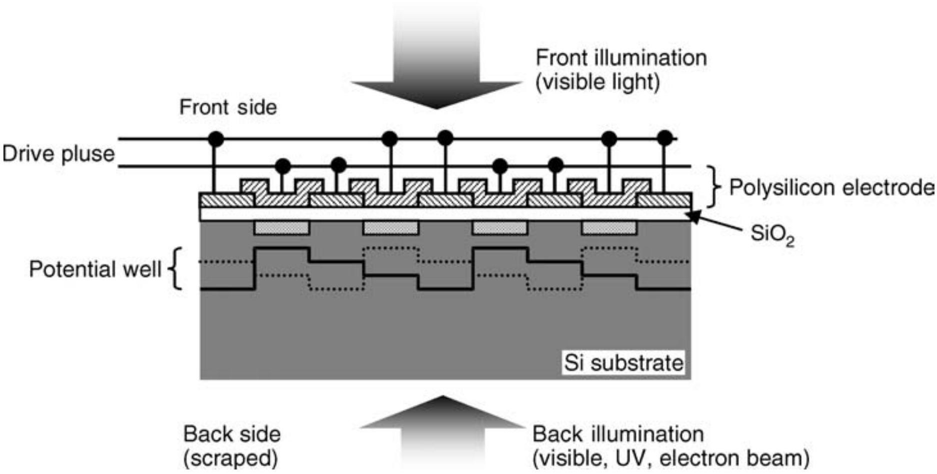
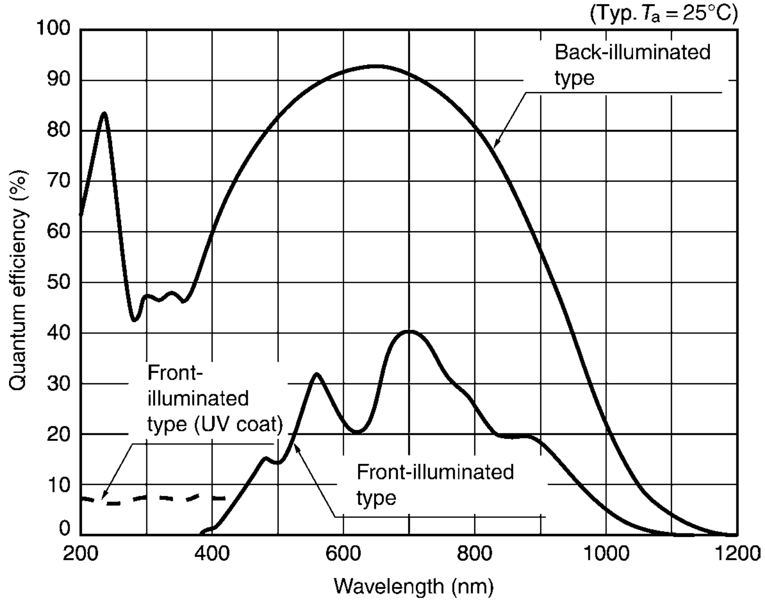

Front illumination
(visible light)
Front side
Drive pluse
Polysilicon electrode
SiO$_2$
Potential well
Si substrate
Back side
(scraped)
Back illumination
(visible, UV, electron beam)

Figure B2.42. Structure of a back-illumination CCD.

The binning operation unique to CCD image sensors is described next. Figure B2.44 shows the principle of the binning operation in the CCD image sensor. In the FFT-CCD, the signal charges are stored in the potential well during the integration time, and when integration time ends the signal charges are stored two-dimensionally. Here, signals can be summed in the line direction (line binning) as shown in the figure, by transferring charges separately to the vertical shift register and horizontal shift register. Besides vertical binning that adds signals perpendicularly in this way, binning can also be performed horizontally.



Hamamatsu Photonics; Image Sensors, Cat. No. KMPD0002E01(2001)

Figure B2.43. Spectral response characteristics of a back-illumination CCD (without window).

Binning allows using a two-dimensional image sensor as a one-dimensional image sensor (line binning) and changing the number of effective channels (horizontal binning).

Spectrophotometric applications of semiconductor image sensors are expanding to fluorescence spectrometry, ICP, Raman spectrometry, and multichannel spectrometry. Imaging applications include fields such as bio-imaging, semiconductor device analysis, and astronomical observation.

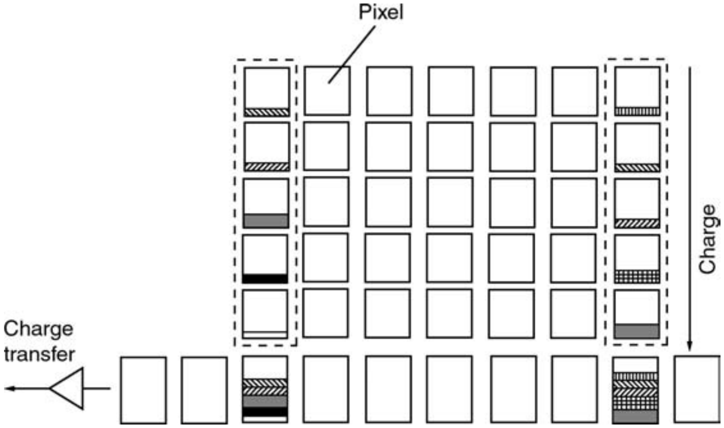


Figure B2.44. Principle of binning operation in the CCD image sensor.

B2.6.3 Hybrid image sensors

Combining electron tube image sensors with semiconductor image sensors allows boosting performance much more than would be possible with a single type of image sensor. Hybrid image sensors with additional functions are now seeing wide practical use.

One type of hybrid image sensor is a high-sensitivity multichannel detector [40] combining an II with a semiconductor linear image sensor. Figure B2.45 shows a high-sensitivity multichannel detector combining an II with a semiconductor linear PD array (512-channel MOS PD array) as well as its signal processing diagram. Light input to the II is converted to photoelectrons by the photocathode, amplified 10^3-10^4 times by the MCP, and converted to an optical image on the phosphor screen. The optical image on the phosphor screen is then photoelectrically converted by the linear image sensor, and the signal from the PD array is output as a video signal through a signal processing circuit. A gain of 10^6 or more can be obtained when a two- or three-stage MCP is used, allowing photon counting [43]. High-speed shutter operation on the subnanosecond level can also be obtained by applying a gate pulse to the II as needed.

The high-sensitivity multichannel detector described here has applications in observation of internal combustion engines, plasma, fluorescence, biochemical emissions, calcium ion density distribution, Raman spectrometry, semiconductor device evaluation, etc. Its use has also become common on production lines, laboratories, and field test measurements.

Figure B2.46 shows the structure of an intensified CCD (ICCD) [44] developed for imaging of low-level light. The ICCD amplifies low-level input light in an II and the output image produced on the phosphor screen is read out by an externally coupled CCD area image sensor. Amplification in an ICCD depends on the number of MCPs incorporated in the II and the applied voltage. Under high-sensitivity conditions the ICCD is even capable of photon counting (photon counting imaging) [45].

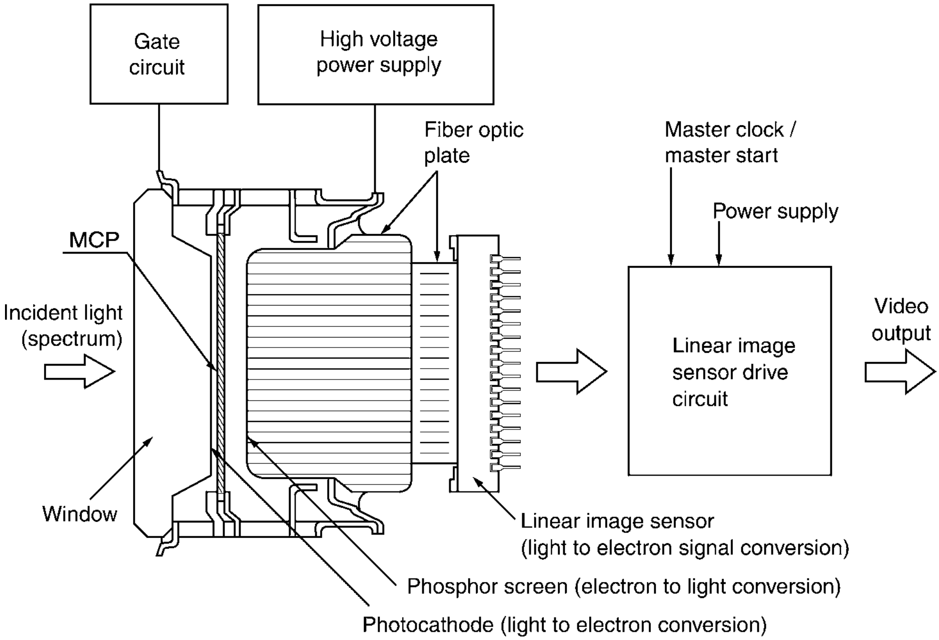

Gate circuit
High voltage power supply
Fiber optic plate
Master clock / master start
Power supply
MCP
Incident light (spectrum)
Linear image sensor drive circuit
Video output
Window
Linear image sensor (light to electron signal conversion)
Phosphor screen (electron to light conversion)
Photocathode (light to electron conversion)

Figure B2.45. Structure of a high-sensitivity multichannel detector.

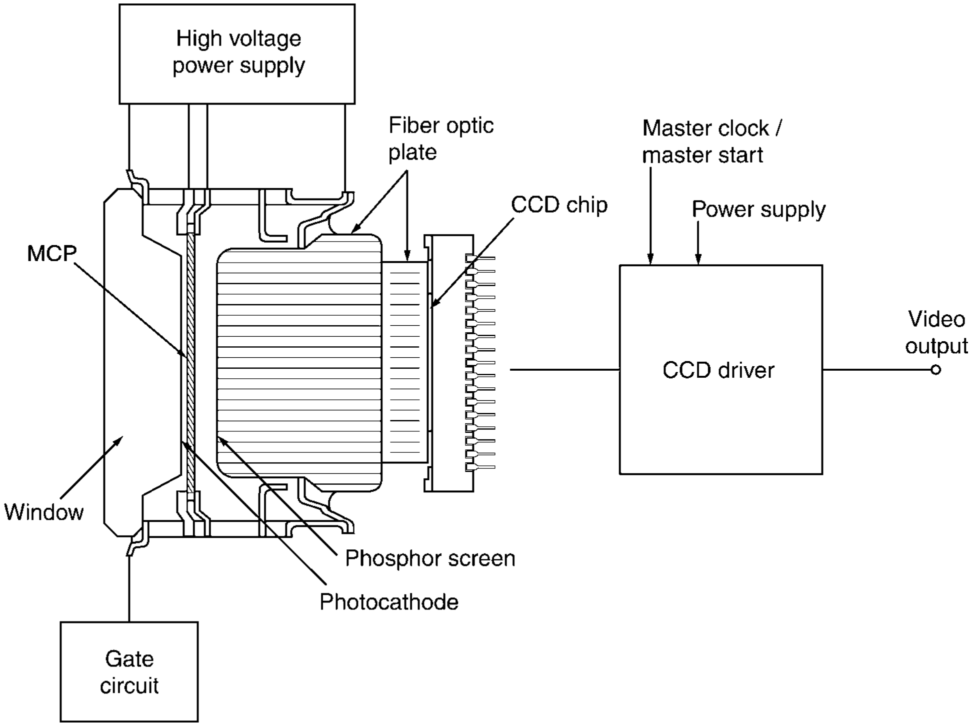


Figure B2.46. Structure of an ICCD.

The ICCD is utilized in imaging applications in the state-of-the-art fields such as combustion analysis in internal combustion engines, imaging of high-speed moving objects, and bioluminescence imaging.

In recent years, a unique imaging device called electron bombardment CCD (EBCCD) [46, 47] has been developed for low-light-level imaging. The EBCCD uses a back-illumination CCD sealed in a photoelectric vacuum tube. Electrons emitted from the photoelectric surface (photocathode) are directly detected by the internal CCD. As shown in figure B2.47, the EBCCD utilizes the gain resulting from electron bombardment multiplication occurring when accelerated electrons lose energy and generate many electron–hole pairs within the CCD. This device has the advantage of a high numerical aperture and low gain fluctuation. The EBCCD yields a gain of 50–1300 times that is nearly equal to the gain obtainable with an ICCD using a single MCP. The EBCCD is mainly utilized in low-level fluorescence imaging under a microscope.

Figure B2.48 is a table comparing image sensor detection sensitivity such as in CCD and ICCD image sensors. The CCD has no amplifying function and is therefore restricted to use in areas of strong light intensity and also has a limited dynamic range due to the charge saturation and noise level. The ICCD takes advantage of the MCP amplification effect to achieve a wide dynamic range and ultra-high sensitivity by changing the number of MCP stages and applied voltage. As mentioned earlier, the II not only delivers ultra-high sensitivity but can also function as a high-speed shutter camera by joint use with an ultra-high-speed gate on the nanosecond to subnanosecond level.

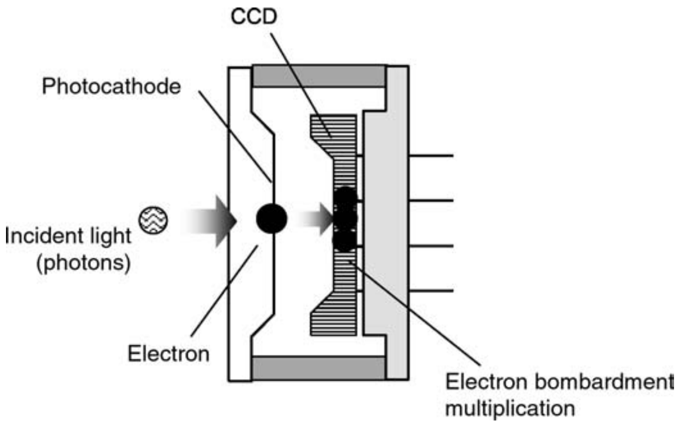

CCD
Photocathode
Incident light (photons)
Electron
Electron bombardment multiplication

Figure B2.47. Structure of an EBCCD.

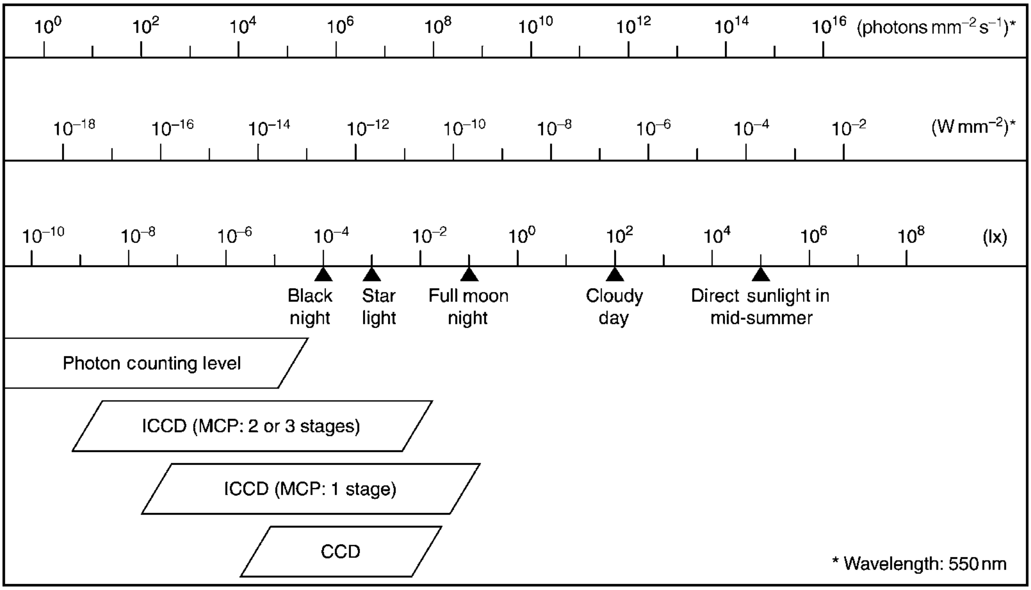


Figure B2.48. Comparison of image sensor detection sensitivity.

B2.7 Future prospects

This chapter has discussed categories, basic principles and various characteristics of photodetectors or optical detectors. Future years will show accelerated progress in applying light to the measurement, machining, control technology, communications, general electronics, and basic science fields. Along with these advances, extensive effort will be poured into developing practical photodetectors with higher sensitivity and time response, greater miniaturization or integration and multichannel detection capabilities.

References

[1] Sommer A H 1980 *Photoemissive Materials* (Malabar, FL: Krieger)
[2] Bube R H 1992 *Photoelectronic Properties of Semiconductors* (Cambridge: Cambridge University Press)
[3] Hamamatsu Photonics K. K. Editorial Committee 1999 *Photomultiplier Tubes—Basics and Applications* 2nd edn (Hamamatsu Japan)
[4] Dereniak E L 1984 *Optical Radiation Detectors* (New York: Wiley)
[5] Driscoll W G 1978 *Handbook of Optics* (New York: McGraw-Hill)
[6] Samson J A R 1967 *Techniques of Vacuum Ultraviolet Spectroscopy* (New York: Wiley)
[7] Knoll G F 2000 *Radiation Detection and Measurement* 3rd edn (New York: Wiley)
[8] Endo T *et al* 1993 *SPIE* **1982** 186
[9] Kosel P B *et al* 1994 *SPIE* **2685** 140
[10] Razeghi M *et al* 1996 *SPIE* **2685** 114
[11] Breskin A 1996 *Nucl. Instrum. Meth. Phys. Res.* **A371** 116
[12] Niigaki M *et al* 1997 *Appl. Phys. Lett.* **71-17** 2493
[13] Keyes R J 1977 *Optical and Infrared Detectors* (Berlin: Springer)
[14] Willardson R K 1977 *Semiconductors and Semimetals* vol 12 (New York: Academic)
[15] Carr J J 1997 *Electronic Circuit Guidebook: Electro-Optics* (Prompt Publications Div. Sams)
[16] Wobschall D 1979 *Circuit Design for Electronic Instrumentation* (New York: McGraw-Hill)
[17] Hamamatsu Photonics K K 2000 *Photon Counting*, TPHO 9001E02
[18] Candy B H 1985 *Rev. Sci. Instrum.* **56** 194
[19] Theodórsson P 1996 *Appl. Radiat. Isot.* **47** 827
[20] Vetokhin S S *et al* 1987 *Sov. J. Opt. Technol.* **54** 754
[21] Pruett H D 1972 *Appl. Opt.* **11** 11 2529
[22] Shapiro S L (ed) 1977 *Ultrashort Light Pulses, Picosecond Techniques and Applications, Topics in Applied Physics vol 18* (Berlin: Springer)
[23] Harris C B, Ippen E P, Mourou G A and Zewail A H (ed) 1990 *Ultrafast Phenomena V* II, *Springer Series in Chemical Physics* vol 53 (Berlin: Springer)
[24] Alfano R R (ed) 1982 *Biological Events Probed by Ultrafast Laser Spectroscopy* (New York: Academic)
[25] Tsuchiya Y 1991 *SPIE* **1599** 244
[26] Yamazaki I *et al* 1985 *Rev. Sci. Instrum.* **1187**
[27] Chou S Y, Liu Y and Fischer P B 1991 *IDEM* **91** 745
[28] Takahashi A *et al* 1994 *Proc. SPIE* **2116** 275
[29] *ITU-T Recommendation*, G671 to be approved in 2002
[30] Tan I H *et al* 1995 *IEEE Photon Technol Lett.* **1477**
[31] NEC Compound Semiconductor Devices, Ltd 2002 *Optical Semiconductor Devices For Fiber Optic Communications Selection Guide* Document No. PX10161EJOIVOPF
Hamamatsu Photonics 2001 Optical communication device KOTH0005E01
[32] Csorba I P 1985 *Image Tubes* (Sams)
[33] Biberman L M 2000 *Electro-Optical Imaging* (SPIE)
[34] Kume H *et al* 1990 *SPIE* **1358** 1444
[35] Thomas S 1990 *SPIE* **1358**
[36] Shionoya S and Yen W M 1999 *Phosphor Handbook* (Boca Raton, FL: CRC Press)
[37] Janesick J R 2001 *Scientific Charge-Coupled Devices* (SPIE)
[38] Séquin C H 1975 *Charge Transfer Devices* (New York: Academic)
[39] Schroder O K 1987 *Advanced MOS Devices* (Wesley)
[40] Sweedler J V *et al* 1989 *Appl. Spectrosc.* **43–46** 953
[41] Theuwissen A J P 1995 *Solid-State Imaging with Charged-Coupled Device* (Dordrecht: Kluwer)
[42] Muramatsu M *et al* 1997 *SPIE* **3019**
[43] Tsuchiya Y, Inuzuka E, Kurono T and Hosoda M 1985 Photon-counting imaging and its application *Adv. Electron. Electron Phys.* **64A** 21
[44] Hirano A *et al* 1993 *Japan. J. Appl. Phys.* **32** 3300
[45] Tsuchiya Y, Inuzuka E, Kurono T and Hosoda M 1985 Photon-counting image acquisition system and its applications *J. Imaging Technol.* **11** 215
[46] Williams G M Jr *et al* 1995 *Proc. SPIE* **2551** 208
[47] Suyama M *et al* 1997 *Proc. SPIE* **3173** 422

B3
Optical fibre devices

Suzanne Lacroix and Xavier Daxhelet

B3.1 Introduction

Maximizing the capacity of the fibre as a transmitting medium is the major challenge met by network operators. Solitonic propagation as well as wavelength division multiplexing (WDM) are among the solutions presently proposed to increase the flow of information propagating in these networks. This progress is made possible thanks to the development of such essential components as the erbium doped fibre amplifier (EDFA) and frequency stabilized laser sources. However, other components to perform all the functions (routing, filtering, dispersion compensation, etc) are not less essential. Different approaches to design and realize the corresponding components are commonly used. The advantage of the all-fibre approach over its competitors (micro- and integrated optics) lies undoubtedly in the fact that all-fibre components are readily integrated to the network without significant splicing loss. In addition, their polarization sensitivity (which would induce loss and dispersion) is intrinsically much smaller than that of their integrated optic counterparts.

The present chapter content is restricted to all-fibre components, their micro-optics and integrated-optics equivalents being treated in chapters A16 and B8, respectively. Unless expressly mentioned, these all-fibre components are made of single-mode fibre. The emphasis is put on the way they function and their intrinsic limitations, leaving out details about fabrication and packaging problems, which are treated elsewhere (see chapter B12). Besides, only passive components are considered. Optical amplifiers and lasers can be found in chapters A18 and B6. Some other more advanced components, such as wavelength converters, are based on non-linear effects (see chapter A2.4). They usually take advantage of the inherent third-order non-linear effects, but second-order effects can also be induced as explained in chapter B14. These components are not treated in the present chapter.

In the following section, we present the basic technologies that are involved. The subsequent section gives concepts that determine the behaviour of these all-fibre components. Example components are presented in the third section.

B3.2 Technologies

Two main technologies are presently used to manufacture all-fibre components:

1. the fusion and tapering technology;
2. the inscription of gratings including both short-period gratings (usually referred to as Bragg gratings) and long-period gratings (LPGs).

B3.2.1 Fusion and tapering

The tapering technique consists of locally heating and stretching a fibre using a micro-torch or a CO_2 laser, thus creating a biconical structure such as that of figure B3.1.

Before tapering, one may laterally fuse two (or more) fibres together so as to create a more complex transverse structure in order to transfer power from one guide to another. Such structures are referred to as couplers.

Irrespective of the transverse structure, the behaviour of the component is largely determined by the slopes of the longitudinal structure. In the case of a tapered single fibre, however, the angle is made small enough everywhere so that only a negligible leakage of power from the fundamental mode as it propagates along the structure is ensured. In such cases, the propagation and, by extension, the taper itself is said to be *adiabatic*, where the fibre transmission is not affected by the tapering process [1]. In contrast, when the slopes are abrupt, such as those of the structure shown in figure B3.1, one may observe large oscillations in the transmitted power, as the fibre is elongated. In addition, for a given elongation, similar oscillations are seen in the transmission as a function of wavelength (figure B3.20) or of the refractive index surrounding the tapered part of the fibre [2].

While adiabaticity is usually required for couplers, non-adiabaticity of tapered single fibres can be used to design a variety of all-fibre spectral filter.

B3.2.2 Gratings

All-fibre short- and long-period gratings are periodically perturbed fibres. The number of periods ranges from ten to several thousands depending on the intensity of the index perturbation (up to 10^{-2}). Both short- and long-period gratings operate according to a resonant coupling effect resulting in a condition relating to the modal wave vectors $\vec{\beta}_1$ and $\vec{\beta}_2$ with the grating one $\vec{\beta}_B$

$$\vec{\beta}_1 = \vec{\beta}_2 + \vec{\beta}_B. \tag{B3.1}$$

Wave vector resonance conditions are shown in figure B3.2 for contra- and codirectional coupling. Short periods $\Lambda = 2\pi/\beta_B$ are of the order of 0.5 μm and long ones 500 μm. As their periods are of the order of half the wavelength in silica, short-period gratings reflect the fundamental mode into itself (through contradirectional coupling) at a given wavelength ($\vec{\beta}_2 = -\vec{\beta}_1$). They also reflect higher-order cladding modes at shorter wavelengths with lesser efficiency, unless especially designed for this purpose. On the other hand, long-period gratings couple, at a given wavelength, the fundamental mode into

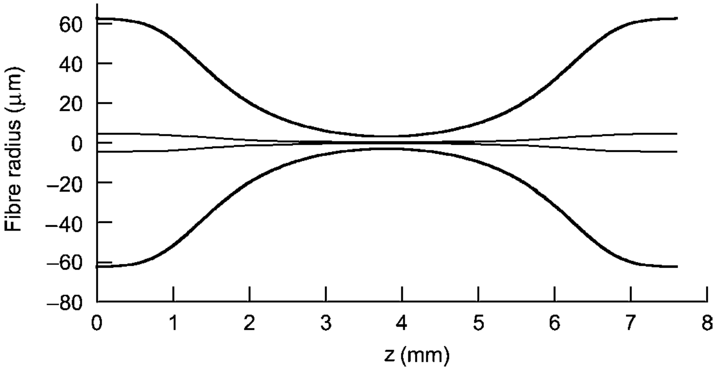


Figure B3.1. Profile of an abruptly tapered fibre.

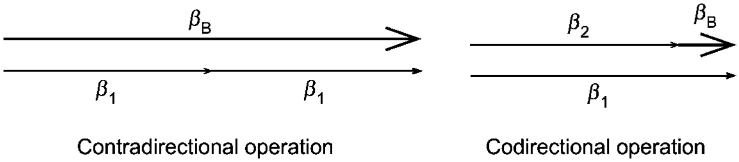


Figure B3.2. Wave vector conservation in fibre gratings. Contradirectional coupling corresponds to large grating wave vectors β_B and thus small grating periods Λ (typically half a μm), while codirectional coupling corresponds to small grating wave vectors β_B and thus large grating periods Λ (several hundred μm).

copropagating higher-order modes (through codirectional coupling), which are eventually absorbed by the fibre jacket. Both types of grating are thus rejection filters in transmission.

Short-period gratings

There are essentially two methods to manufacture this type of grating. Both are, however, based on photosensitivity—a stable refractive index change of the order of a few 10^{-4} up to 10^{-2} that the Ge-doped core of the fibre experiences, when exposed to near UV-radiation.

1. The first method consists of exposing the fibre to an interference field created by two locally planar waves of a UV-laser such as doubled Ar^+, dye lasers or copper vapour lasers. The period Λ is then given by

$$\Lambda = \frac{\lambda_{\mathrm{UV}}}{2\sin(\theta/2)} \tag{B3.2}$$

where λ_{UV} is the writing wavelength and θ the angle between the interfering waves. This method referred to as the holographic method is very flexible because any periodicity may be realized. A number of practical implementations have been designed [3, 4]. However, as for any holographic setup, stability is an issue. It is thus easier to implement it in a pulsed regime.

2. The alternative method uses a phase mask as schematized in figure B3.3. It is a phase grating having a periodicity Λ_m, twice that of the grating to be written in the fibre ($\Lambda_m = 2\Lambda$). Its depth e and refractive index n (at λ_{UV}) are related by $2(n-1)e = \lambda_{\mathrm{UV}}$ ensuring that the zero order is not transmitted. Waves from orders $+1$ and -1 thus interfere in a similar manner to that described above for the holographic method.

The main advantage of the phase mask method over the holographic one is that the UV-source may have a limited spatial and temporal coherence. This is the case of excimer lasers (KrF lasing at $\lambda_{\mathrm{UV}} = 248$ nm and ArF at $\lambda_{\mathrm{UV}} = 193$ nm), which are frequently used for fibre grating inscriptions. The fibre and the mask are then separated by only a few micrometres. The disadvantage of this method is that the Bragg period is determined by that of the phase mask, which is an expensive component.

This drawback may, however, be overcome by a combination of both methods in set-ups such as the Talbot interferometer in which the phase mask is used as a splitting plate and the fringe periodicity is varied through a mirror angle [3, 4].

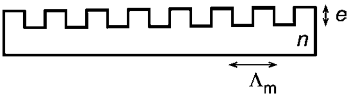


Figure B3.3. Phase mask.

Long-period gratings

As their period is long ($\Lambda = 300-800\,\mu\text{m}$), this type of grating usually does not require sophisticated set-ups such as those of their short-period counterparts. Most of the time they use step-by-step methods.

1. The fibre may be irradiated by a UV-source just as in the short-period case. One may use the step-by-step method or an amplitude mask.
2. CO_2-laser irradiation ($\lambda = 10, 6\,\mu\text{m}$) also induces index changes of the same order of magnitude, but the mechanism responsible for this change, although not completely identified, is certainly different as the perturbation persists at higher temperatures than photosensitivity [5].
3. Larger index changes (up to a few 10^{-3}) may also be induced by an electrical discharge. The grating is also stable at high temperatures and its inscription is clearly related to the fibre inner stresses [5].
4. Femtosecond intense Ti:sapphire lasers are also known to induce large refractive index changes that may be used to manufacture long-period gratings.
5. Finally, one may manufacture LPGs by periodical tapering of the fibre.

All these fabrication techniques (fusion-tapering and grating inscription) may be viewed from a theoretical viewpoint as perturbations to the longitudinal invariance of the fibre structure. These perturbations in turn induce coupling between the modes and thus power the transfer from the fundamental mode to others. These effects are described by the coupled mode formalism, which is now presented.

B3.3 Coupled mode theory

Except for the propagation factor $e^{i\beta_j z}$ along the guide, modes of an unperturbed guide, identified by their indices j, are z-invariant solutions of Maxwell equations. Any solution, i.e. any electromagnetic field propagating along the guide, is a superposition of these modes with amplitudes a_j.

When the guide is perturbed, these modes are no longer solutions and the optical power may transfer from one mode to another. This is described by permitting the modal amplitudes a_j to vary with z.

Whatever the perturbation, periodical or not, resonant or not, between co- or contradirectional modes, it always consists of a variation of the refractive index, i.e. of the polarization vector. Coupled mode equations may be demonstrated from Maxwell equations by expanding the electromagnetic field on a modal basis [6]. These equations may be written as

$$\frac{da_j}{dz} - i\beta_j a_j = i\sum_{\ell} C_{j\ell}(z)a_{\ell} \tag{B3.3}$$

where a_j and β_j are, respectively, the amplitude and the propagation constant of mode j and $C_{j\ell}(z)$ the coupling coefficient between mode j and mode ℓ. This coupling coefficient may be real or complex depending on the waveguide perturbation. As the set of all the modes (guided as well as radiation modes) of an unperturbed guide constitutes a basis for the expansion of any electromagnetic field, it is natural to choose a_j to be amplitudes of the normal modes of a guide. The choice of the guide used for the basis depends on the problem as will be shown in the following section.

B3.3.1 Ideal mode coupling

As shown in figure B3.4, whenever the perturbation is a slight departure from a z-invariant guide, referred to as the *ideal guide*, the natural basis is the set of modes of this ideal guide. One also refers to this expansion as the *ideal mode* expansion. The coupling coefficient in the scalar approximation [6] is shown to be

$$C_{j\ell} = \frac{k^2}{2\sqrt{|\beta_j\beta_\ell|}} \int_{A_\infty} (n^2 - \bar{n}^2)\hat{\psi}_j^* \hat{\psi}_\ell \, \mathrm{d}A \tag{B3.4}$$

where $n(x, y)$ is the perturbed guide index profile, $\bar{n}(x, y)$ the unperturbed one, A_∞ the guide cross-section area $\hat{\psi}_j$ and $\hat{\psi}_\ell$ the normalized *scalar* normal fields obeying

$$\int_{A_\infty} \hat{\psi}_j^* \hat{\psi}_\ell \, \mathrm{d}A = \delta_{j\ell} \tag{B3.5}$$

with $\delta_{j\ell}$ the Krönecker symbol.

B3.3.2 Local mode coupling

Whenever, as shown in figure B3.5, the guide varies slowly along z, one can define, in each z cross-section plane, a guide which locally coincides with the perturbed guide. Then, the so-called *local mode* expansion is performed. In this case, as opposed to the former one, the basis modes depend on z in particular through their propagation constants $\beta_j(z)$. The coupling coefficient, in the scalar approximation, is shown when one neglects longitudinal variations of the modal field

$$C_{j\ell} = -\mathrm{i}\frac{k^2}{2\sqrt{|\beta_j\beta_\ell|}} \frac{1}{\beta_j - \beta_\ell} \int_{A_\infty} \frac{\partial n^2}{\partial z} \hat{\psi}_j^* \hat{\psi}_\ell \, \mathrm{d}A. \tag{B3.6}$$

B3.3.3 Individual guide coupling

In the case of couplers made of single-mode waveguides, as shown in figure B3.6, the amplitudes a_j and a_ℓ of equation B3.3 are the fundamental mode amplitudes of the unperturbed guides (i.e. without the other guides). Although obviously not constituting a basis, these modes serve as an expansion set, which permits us to describe the power exchange between the guides quite accurately. The coupling coefficient takes the same form as in equation (B3.4) although, here, $\hat{\psi}_j$ and $\hat{\psi}_\ell$ are the normalized modal fields of individual guides j and ℓ; the unperturbed index profile $\bar{n}(x, y)$ is that of the individual guide supposed alone and $n(x, y)$ the index profile of the whole coupler.

Note, however, that in the case of a fused coupler, single-mode guides are not defined for any cross-section so that one must abandon this concept and use supermodes instead, as described in more detail in section B3.3.1.

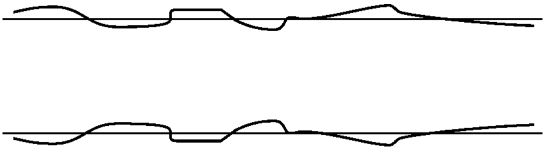


Figure B3.4. Perturbation of an ideal guide.

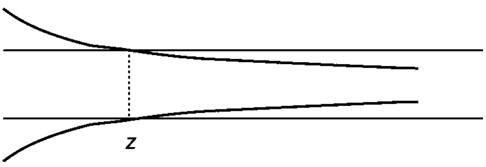


Figure B3.5. Perturbation of local guide.

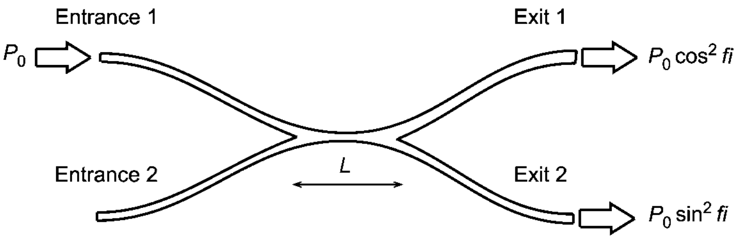


Figure B3.6. Diagram of a 2 × 2 coupler.

B3.3.4 Beating and coupling lengths

The beating and coupling lengths are the key parameters to predict the behaviour of components.

Let $C_{j\ell}$ be the coupling coefficient between modes j and ℓ. Whenever the coupling coefficients do not depend on z, solutions of the coupled mode equation (B3.3) are sinusoidal functions. The coupling length L_{C} is defined as the length for which a complete *power* transfer cycle takes place (transfer from mode j to mode ℓ and back to mode j). One thus defines

$$L_{\mathrm{C}} = \frac{2\pi}{|C_{j\ell}|}. \tag{B3.7}$$

The beating length z_{b} between these two modes is defined as the length along which the modes accumulate a 2π phase difference. One thus has

$$z_{\mathrm{b}} = \frac{2\pi}{|\beta_j - \beta_\ell|}. \tag{B3.8}$$

These length scales permit us to define a criterion to discriminate between slowly varying adiabatic structures and abruptly varying ones for which modes are coupled. A rule of the thumb is given by the condition

$$L_{\mathrm{C}} = z_{\mathrm{b}}. \tag{B3.9}$$

1. If the coupling length of two given modes is larger than their beating length $L_{\mathrm{C}} > z_{\mathrm{b}}$, the modes accumulate a phase difference through the beating phenomenon without an exchange of power. The process may be considered as an adiabatic one.
2. If the coupling length of two given modes is smaller than their beating length $L_{\mathrm{C}} < z_{\mathrm{b}}$, the modes exchange power through the coupling process over such distances that their propagation phase difference is negligible. The process is then non-adiabatic.

B3.4 All-fibre passive components

Passive components are defined as components that do not necessitate an external source of energy, as opposed to lasers, amplifiers and other active components. Most of the time, all-fibre passive components are based on the technologies such as fusion-tapering and grating inscription described earlier. The fabrication parameters of an individual structure, tapered fibre coupler or grating, may be varied to obtain a prescribed behaviour. In addition, several individual components may be concatenated, cascaded or mixed to accomplish more complex functionalities. The following sections give the essential basic concepts to understand the principles underlying the operation of all these components and describe some of the example components based on these principles.

These components are characterized by several parameters. Essential parameters are IL, CD, PDL and DGD.

1. The *insertion loss* (IL) is the transmittance in dB, defined by $\mathrm{IL} = 10 \log T$.
2. The *chromatic dispersion* (CD) is the pulse broadening per unit wavelength and length of propagation, resulting from both material and waveguide contributions. For fibres, it is expressed in $\mathrm{ps\,nm^{-1}\,km^{-1}}$ and, for step index profiles, the main contribution comes from the material CD, i.e. the dependence of the index of silica on wavelength. In components, it is expressed in $\mathrm{ps\,nm^{-1}}$ and is usually small except for gratings, which may be designed to compensate the dispersion accumulated over several kilometres of fibre.
3. The *polarization dependent loss* (PDL) is due to the component birefringence. This birefringence results an IL depending on polarization. The PDL parameter is defined by the difference of the ILs for the best and worst polarization states and is thus expressed in dB.
4. The *differential group delay* (DGD) characterizes the variation of the group delay with polarization. It is defined by the difference between group delays for the fastest and slowest polarization states, and is usually expressed in ps. Note that the length of fibre is characterized by its *polarization mode dispersion* (PMD) expressed in $\mathrm{ps\,km^{-1/2}}$, as the DGD accumulates statistically as the square root of distance of propagation.

B3.4.1 Splitters and combiners

All-fibre splitters and combiners are based on couplers which are, most of the time, made by the fusion and tapering technique. Although a coupler may consist of an arbitrary number of fused possibly different fibres, most of them are 2×2 couplers made of identical fibres. Two geometric parameters determine their behaviour:

- Their degree of fusion f that may theoretically vary from zero (tangent fibres) to one (completely fused fibres resulting in a circular cross-section) [7].
- Their longitudinal profile that is described by the variation of the inverse taper ratio (ITR-) parameter along z. In a first approximation, the tapering process is assumed to preserve the structure respective dimensions, so that one can characterize the transverse reduction by the ratio of a given length (e.g. the core radius or the coupler width) measured after and before the tapering process. This is the definition of the ITR-parameter.

Principle of operation

Concept of supermodes

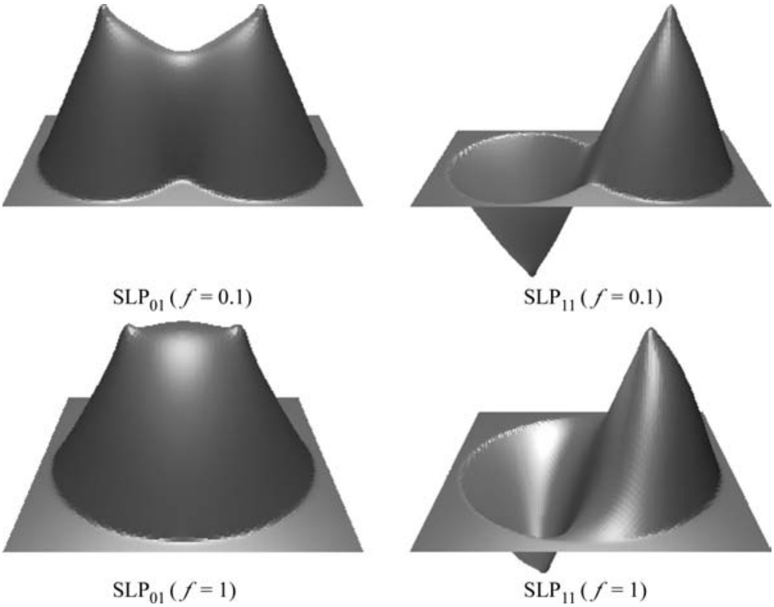
As a fused and tapered coupler is not longitudinally invariant, one must work with the *local modes* (or in other words, *local supermodes*) of the fused structure. As explained later, in identical fibre couplers, the power exchange from the main to the secondary branch, and vice versa, takes place via the *beating* phenomenon between supermodes. Supermode *coupling*, which would arise from very abrupt slopes, is undesirable as it would create loss. Symmetric fused couplers are thus, most of the time, adiabatic structures.

1. At the coupler entrance, the fibres are not tapered: modes of individual guides are confined in the cores, and one can consider that the supermodes are combinations of individual guide modes. The two supermodes of a symmetrical 2 × 2 coupler, i.e. made of two identical fibres, are

$$|\Psi_+\rangle = \frac{|\psi_1\rangle + |\psi_2\rangle}{\sqrt{2}} \quad \text{and} \quad |\Psi_-\rangle = \frac{|\psi_1\rangle - |\psi_2\rangle}{\sqrt{2}} \tag{B3.10}$$

where the Dirac notation $|\psi_1\rangle$ and $|\psi_2\rangle$ is used to describe the fields of the fundamental (scalar) LP_{01} modes of guides 1 and 2. The supermodes corresponding to $|\Psi_+\rangle$ and $|\Psi_-\rangle$ are, respectively, called SLP_{01} and SLP_{11} after their circular two-layer fibre counterparts as can be seen in figure B3.7 for degree of fusion $f = 1$. Whenever one excites $|\psi_1\rangle$, the fundamental mode of fibre 1 at the entrance of the coupler, one actually excites a superposition of both supermodes, which can be written as

$$|\psi_1\rangle = \frac{|\Psi_+\rangle + |\Psi_-\rangle}{\sqrt{2}} \quad \text{and} \quad |\psi_2\rangle = \frac{|\Psi_+\rangle - |\Psi_-\rangle}{\sqrt{2}}. \tag{B3.11}$$



SLP_{01} ($f = 0.1$)

SLP_{11} ($f = 0.1$)

SLP_{01} ($f = 1$)

SLP_{11} ($f = 1$)

Figure B3.7. Fields of the first two supermodes of a 2 × 2 coupler for two different values of the degree of fusion and for an ITR = 0.1.

All these modes are degenerate, i.e. the supermode propagation constants, β_+ and β_-, are identical to that of the each individual guide fundamental mode, β ($\beta = \beta_1 = \beta_2 = \beta_+ = \beta_-$).

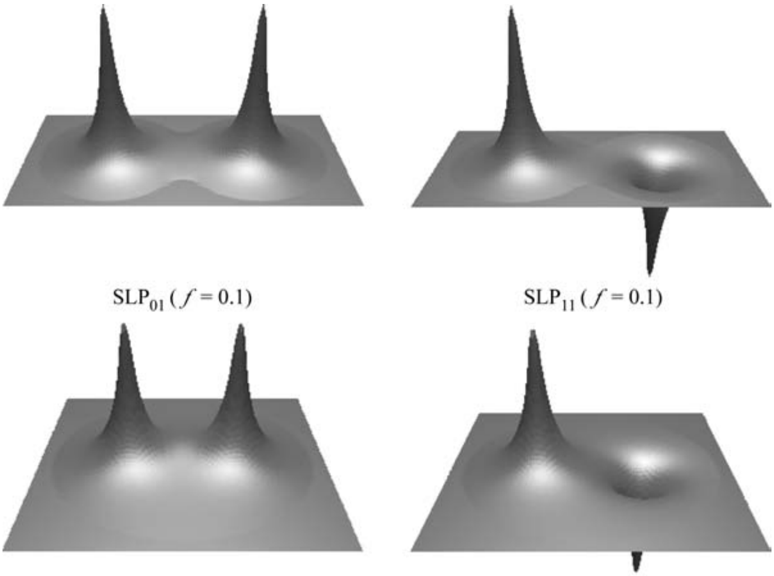
Figure B3.8 shows, as an example, supermode fields for two different degrees of fusion ($f = 0.1$ and 1) in the case of a core guiding structure. For a coupler made of standard telecommunication fibres, supermodes may be approximated by superpositions of individual fibre modes as long as ITR > 0.4.

2. In the central region of the coupler, the individual guides lose their identity and the cores in their guiding roles. Only supermodes keep a physical meaning. Figure B3.7 shows the supermodes for the same degrees of fusion as in figure B3.8, but for ITR $= 0.1$.

As shown in figures B3.9 and B3.10 for two different degrees of fusion, the effective indices $n_{\mathrm{eff}} = n_\pm$ (and thus propagation constants $\beta_\pm = 2\pi n_\pm/\lambda$) of the first two supermodes (SLP_{01} and SLP_{11}) split apart in the central region of the coupler (for which ITR < 0.4). Supermodes accumulate a phase difference as they propagate along the coupler structure. In other words, it is the supermode *beating* phenomenon which governs the power exchange process.

3. At the coupler exit, the situation resembles that of the entrance, i.e. the guides are sufficiently separated so that the supermodes are again superpositions of individual guide modes and are consequently degenerated. It is thus the phase difference accumulated in the central region that determines the power splitting ratio at the exit of the coupler.

Expressions of the individual modes (equation (B3.11)) show that whenever the supermodes are in phase, the power is recovered in branch 1 and, whenever they are out of phase, it is recovered in branch 2. A 50/50 splitting ratio—which corresponds to the so-called 3 dB coupler—occurs whenever the supermodes exit the coupler with a $\pi/2$ phase difference.



SLP_{01} ($f = 0.1$) SLP_{11} ($f = 0.1$)

SLP_{01} ($f = 1$) SLP_{11} ($f = 1$)

Figure B3.8. Fields of the first two supermodes of a 2 × 2 coupler for two different values of the degree of fusion and for an ITR $= 0.5$.

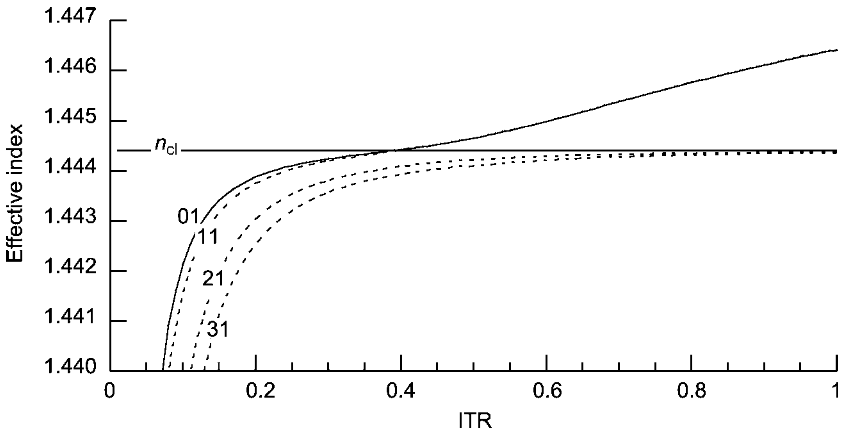


Figure B3.9. Effective indices of the first two symmetric SLP_{01} and SLP_{21} supermodes and of the first two antisymmetric SLP_{11} and SLP_{31} supermodes as functions of ITR at wavelength $\lambda = 1550\,\text{nm}$ for a slightly fused coupler ($f = 0.1$).

Birefringence

As detailed in the following section, a coupler is characterized by its transfer matrix, which relates individual guide amplitudes at the exit of the coupler to the entrance ones. This matrix depends on the accumulated supermode phase difference and thus on their local propagation constants $\beta_+(z)$ and $\beta_-(z)$.

As the transverse cross-section of a coupler does not show the circular symmetry (even though $f = 1$, the cores break off the symmetry), two propagation constants must be attributed to each supermode $SLP_{\ell m}$ corresponding to both polarizations along the symmetry axes. As a result, two transfer matrices must be defined, one for each polarization. Thus, from a general viewpoint, couplers are birefringent and their transmissions are polarization dependent. The less fused the coupler, the more obvious the departure from circular symmetry. One then understands that the birefringence is larger

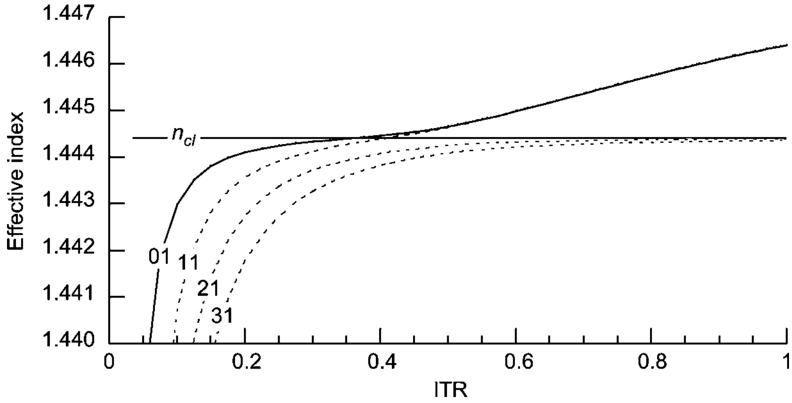


Figure B3.10. Effective indices of the first two symmetric SLP_{01} and SLP_{21} supermodes and of the first two antisymmetric SLP_{11} and SLP_{31} supermodes as functions of ITR at wavelength $\lambda = 1550\,\text{nm}$ for a strongly fused coupler ($f = 1$).

for slightly fused couplers than for those approaching $f = 1$. If the birefringence is negligible, the scalar approximation is sufficient for the calculations. Otherwise, polarization corrections (see chapter 13 of [6]) or exact (i.e. vectorial) calculations must be made.

Transfer matrices

The coupler transmissions as functions of elongation and wavelength greatly depend on the fabrication parameters, degree of fusion and longitudinal profile. These fabrication parameters determine the transfer matrix, which relates individual guide amplitudes at the exit of the coupler to the entrance ones. For a coupler of length L, one has [7]

$$\mathbf{M}_{2\times 2}(\alpha) = \mathrm{e}^{\mathrm{i}\bar{\alpha}} \begin{bmatrix} \cos\alpha & \mathrm{i}\sin\alpha \\ \mathrm{i}\sin\alpha & \cos\alpha \end{bmatrix} \tag{B3.12}$$

with the average propagation phase

$$\bar{\alpha} = \int_0^L \bar{\beta}(z)\,\mathrm{d}z, \qquad \bar{\beta}(z) = \frac{\beta_+(z) + \beta_-(z)}{2} \tag{B3.13}$$

and the supermode accumulated phase difference

$$\alpha = \int_0^L \frac{\beta_+(z) - \beta_-(z)}{2}\,\mathrm{d}z. \tag{B3.14}$$

Here $\beta_+(z)$ and $\beta_-(z)$ are the propagation constants of the local fundamental supermode SLP_{01} and of the first antisymmetric supermode SLP_{11}, respectively. The parameter 2α is the phase difference accumulated by these two supermodes along the coupler.

From the transfer matrix, one may, for a given entrance condition, calculate the transmitted power in both branches. For example, excitation in branch 1 corresponds to a column vector

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and, as a result, in a power transmission in the same branch

$$T_1 = \cos^2\alpha = \frac{1}{2}[1 + \cos 2\alpha] \tag{B3.15}$$

and, in the secondary branch

$$T_2 = \sin^2\alpha = \frac{1}{2}[1 - \cos 2\alpha]. \tag{B3.16}$$

Whenever the modes are in phase ($2\alpha = 0 + 2p\pi$ with p an integer) at $z = L$, the transmission is maximum in the main branch ($T_1 = 1$). Whenever they are out of phase ($2\alpha = \pi + 2p\pi$) at $z = L$, the transmission is minimum ($T_1 = 0$), corresponding to a complete power transfer in the secondary branch ($T_2 = 1$). This power exchange for a strongly fused 2 × 2 coupler is shown in figure B3.11, which displays an experimental response as well as the theoretical one predicted from the fabrication parameters. The polarization effects are clearly visible in this example recording. From a more general viewpoint, the overall transmission is a superposition of both the polarization transmissions. For a polarization entering at 45° from the cross-section symmetry axes x and y, the transmission in the main branch may be written as

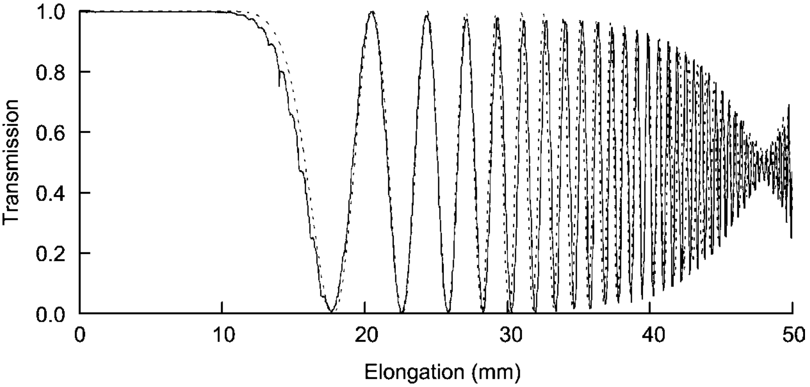


Figure B3.11. Recording (plain line) of the main branch transmission of a strongly fused 2 × 2 coupler as a function of elongation at a given wavelength ($\lambda = 1550\,\text{nm}$). The transmission in the secondary branch, if the losses are negligible, is the complement. The dashed line is the corresponding numerical calculation. Loss of contrast is due to the coupler birefringence.

$$T_1 = \frac{\mathrm{T}_{1x} + \mathrm{T}_{1y}}{2} = \frac{\cos^2\alpha_x + \cos^2\alpha_y}{2} = \frac{1}{2}\left\{\{1 + \cos(\alpha_x - \alpha_y)\cos[2(\alpha_x + \alpha_y)]\right\} \tag{B3.17}$$

with α_x and α_y the supermode accumulated phase difference for polarization x and y, respectively. In figure B3.11, the birefringence appears as an overmodulation of the power exchange with nodes, whenever both polarization transmissions are out of phase ($2\alpha_x - 2\alpha_y = \pi \pm 2m\pi$). In the first power exchange cycles, the birefringence effect is negligible, which permits us to manufacture 3 dB polarization-independent couplers. In contrast, slightly fused couplers present strong birefringence [7].

Oscillations visible in figure B3.11 as a function of elongation also occur as a function of wavelength, which confer to couplers spectral filtering or multiplexing applications as shown in the following sections.

Components

Power splitters

The simplest component to be conceived is the power splitter. According to the figure B3.11 recording, any polarization-independent splitting ratio may be obtained for a relatively short elongation. They are, however, wavelength dependent, which may be undesirable for a number of applications.

Wavelength-independent splitters may be realized using 2 × 2 dissymmetric couplers. As opposed to the symmetric coupler, modes of individual guides of a dissymmetric coupler are not degenerated. When the individual guides are sufficiently separated, at the coupler entrance and exit, the supermodes are the individual guide modes. Quite paradoxically, an adiabatic dissymmetrical coupler would not experience any power exchange. It is thus a supermode *coupling* process (as opposed to a *beating* process) which governs the power transfer in a dissymmetric coupler. This power transfer may be complete or do not depend, for a given asymmetry, on the slopes of the longitudinal structure. The calculation of the adiabaticity criterion as a function of the ITR-parameter is then critical to design a coupler with a prescribed response [8].

An alternative technique is by concatenating two 3 dB couplers in a Mach–Zehnder (MZ) arrangement with a π phase difference between the two branches [9].

Both solutions are used in practice. Any coupling ratio is attainable with these techniques as exemplified in figure B3.12.

Polarization splitters/combiners and depolarizers

Although considered a nuisance for power splitters, one may take advantage of the intrinsic polarization dependence of couplers, to realize polarization splitters—the all-fibre equivalent of polarizing cubes. These all-fibre polarizers consist of couplers elongated until the polarization node (obtained for an elongation of about 47 mm in figure B3.11) is attained. Used in the reverse configuration, they serve as polarization combiners. These polarization splitters/combiners, however, suffer from being narrowband. A wideband alternative consists in an all-fibre MZ structure with a short length (half the beatlength) of a polarization maintaining fibre in one branch. An example response of a wideband polarization splitter/combiner is shown in figure B3.13. These components are particularly useful as polarization pump combiners in Raman amplifiers. For a given power of the pump sources, they allow us to double the pump power at a given wavelength and reduce the impact of the inherent polarization dependence of the Raman gain.

An alternative solution to obtain a polarization-independent Raman amplifier is to depolarize the pump source. This can be done with the help of a depolarizer that consists of an unbalanced interferometer. The arm length difference is longer than the polarization coherence length (which coincides with the usual light source coherence length) so as to uncorrelate the two polarizations. Different designs of depolarizers may be found in [10]. The highest-performance device takes advantage of the MZ polarization combiner (described in the previous section) combined with a fibre ring delay line working in a non-interferometric operation, i.e. with a loop length much longer than the coherence length of the light source. This type of device is characterized by its degree of polarization (DOP) defined as follows

$$\mathrm{DOP} = \sqrt{1 - 4\left[\frac{I_x I_y}{(I_x + I_y)^2}\right](1 - |g_{xy}|^2)} \tag{B3.18}$$

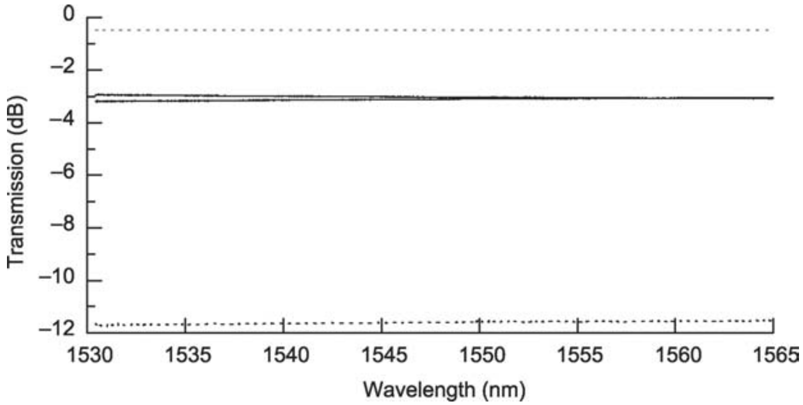


Figure B3.12. Experimental transmissions in dB (also referred to as IL) in both branches of wavelength-independent fused fibre 2 × 2 couplers as a function of wavelength. The plain lines are the transmission of a 3 dB wavelength-independent coupler while the dotted lines are those of a coupler with a large coupling ratio (around 12 dB) (by courtesy of ITF Optical Technologies).

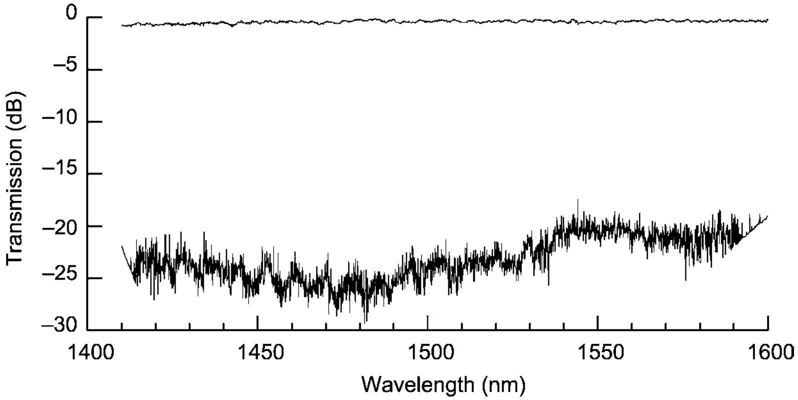


Figure B3.13. Experimental transmissions in one branch for both polarizations of a wideband polarization pump combiner as a function of wavelength. Transmissions in the second branch are similar with polarizations exchanged (by courtesy of ITF Optical Technologies).

where I_x and I_y are the autocorrelation functions for perpendicular polarizations and g_{xy} the cross-correlation normalized coefficient.

An example response of such a depolarizer is shown in figure B3.14. It features a DOP of the order of 10% over a 120 nm spectral band for any temperature between 0 and 70°C. This type of device is also of use for test and measurement applications.

Wavelength splitters/combiners

Fused couplers present an almost inherent sinusoidal wavelength dependence, with a period that decreases when the elongation increases. This wavelength dependence is the basis of several WDM applications. However, their intrinsic properties limit their application.

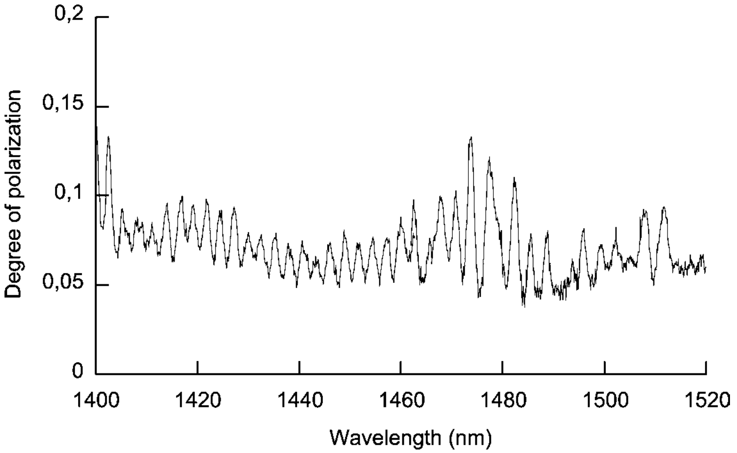


Figure B3.14. DOP as a function of wavelength of a wideband depolarizer. The ripples are due to imperfect alignment of the components in the MZ polarization combiner (by courtesy of ITF Optical Technologies).

1. Even for a relatively large wavelength separation, the inherent sinusoidal spectral response might be a limitation. This is the case for a 1480 nm pump–1550 nm signal multiplexer, useful for Er-doped amplifiers/lasers.
2. The shortest wavelength period attainable is limited by the polarization dependence of the couplers as mentioned earlier. They are consequently not suitable to realize dense wavelength division multiplexing (DWDM) interleavers able to de/multiplex channels as close as 1 nm or less.

In both the cases, alternative solutions may be found by concatenating couplers in MZ arrangements.

1. Figure B3.15 shows, as an example, the spectral response of an MZ 1480–1550 nm WDM allowing a large band around 1550 nm.
2. Dense interleavers are also usually made of an unbalanced MZ structure. Simple MZ consists of two concatenated 3 dB couplers, but more sophisticated non-sinusoidal responses may be obtained by concatenating several MZ structures. Figure B3.16 shows an example of spectral response of such a device. These interleavers are then cascaded in a tree arrangement to effectively separate (or combine) the channels as shown in figure B3.17. Challenges are to make them temperature insensitive with a high isolation, and to respect, over a large spectral range, the ITU grid which is characterized by constant frequency intervals.

Single couplers are nevertheless useful, among other things, to realize de/multiplexers for well-separated signal channels (coarse WDM such as that shown in figure B3.18), pump–signal multiplexers and pump combiners for Er-doped and Raman amplifiers.

Mode splitters/combiners

When made of two-mode fibres, fused couplers may be designed to separate the modes, LP_{01} staying in the main branch while LP_{11} is transferred to the secondary branch [11]. This modal splitter in

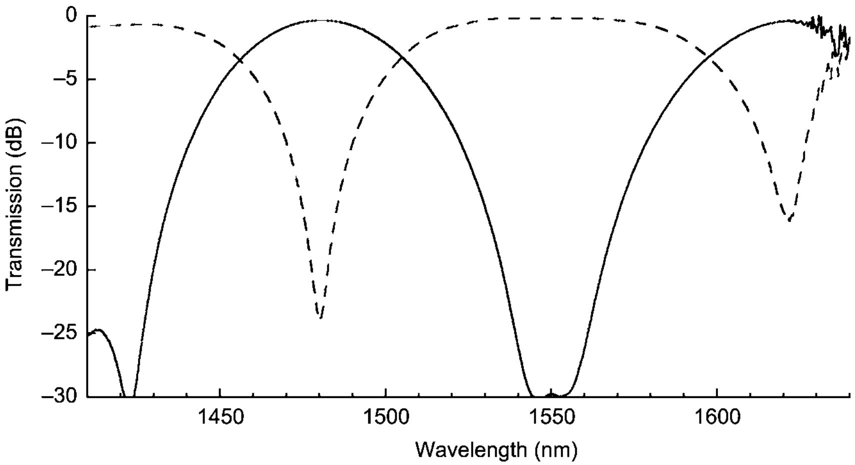


Figure B3.15. Spectral response of a pump–signal WDM combiner for Er-doped amplifiers/lasers (by courtesy of ITF Optical Technologies).

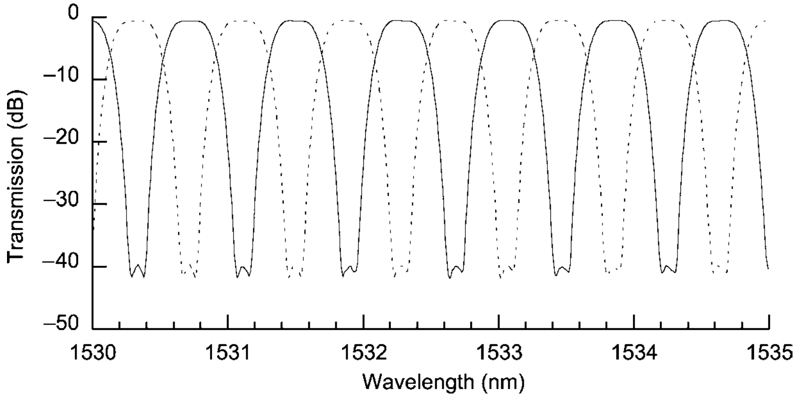


Figure B3.16. Spectral response of a DWDM interleaver with a wavelength spacing of 0.4 nm (50 GHz) (by courtesy of ITF Optical Technologies).

combination with a reflecting modal converter Bragg grating opens up the possibility to get rid of inherently lossy micro-optic circulators.

B3.4.2 Spectral filters and dispersion compensators

Principles of operation

Most of these components, as opposed to couplers, have only one entrance and one exit branch. Some are based on the tapering technology, others on the grating inscription technology. Combination of both the technologies, as in the optical add and drop multiplexers (OADM) described later, may also be used. The tapering technology takes advantage of the sinusoidal wavelength responses of tapered fibres while

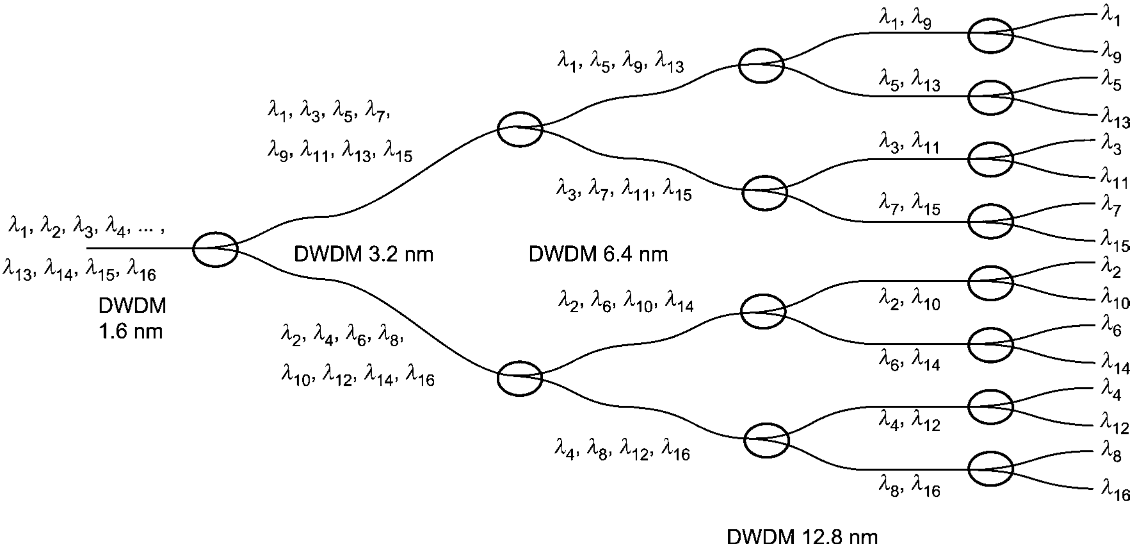

$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots,$
$\lambda_{13}, \lambda_{14}, \lambda_{15}, \lambda_{16}$
DWDM
1.6 nm
$\lambda_1, \lambda_3, \lambda_5, \lambda_7,$
$\lambda_9, \lambda_{11}, \lambda_{13}, \lambda_{15}$
DWDM 3.2 nm
$\lambda_2, \lambda_4, \lambda_6, \lambda_8,$
$\lambda_{10}, \lambda_{12}, \lambda_{14}, \lambda_{16}$
$\lambda_1, \lambda_5, \lambda_9, \lambda_{13}$
$\lambda_3, \lambda_7, \lambda_{11}, \lambda_{15}$
DWDM 6.4 nm
$\lambda_2, \lambda_6, \lambda_{10}, \lambda_{14}$
$\lambda_4, \lambda_8, \lambda_{12}, \lambda_{16}$
λ_1, λ_9
λ_5, λ_{13}
λ_3, λ_{11}
λ_7, λ_{15}
λ_2, λ_{10}
λ_6, λ_{14}
λ_4, λ_{12}
λ_8, λ_{16}
λ_1
λ_9
λ_5
λ_{13}
λ_3
λ_{11}
λ_7
λ_{15}
λ_2
λ_{10}
λ_6
λ_{14}
λ_4
λ_{12}
λ_8
λ_{16}
DWDM 12.8 nm

Figure B3.17. Cascade of $2^n - 1$ WDM splitters to de/multiplex 2^n channels (example wavelength spacing 1.6 nm).

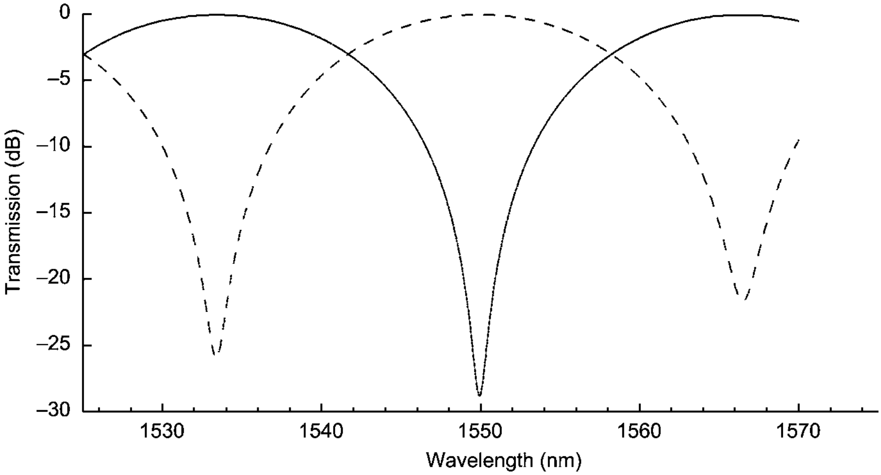


Figure B3.18. Spectral response of a coarse WDM. Degradation due to polarization dependence is visible (by courtesy of ITF Optical Technologies).

gratings are resonant devices: they couple, at a given wavelength λ, two modes of propagation constants β_1 and β_2 through the matching condition of equation (B3.1)

$$(\beta_1 - \beta_2)\Lambda = 2\pi \tag{B3.19}$$

where Λ is the grating period. Mode 1 is usually the fundamental core mode, while mode 2 is, in the case of a short-period grating, the counterpropagating fundamental mode ($\beta_2 = -\beta_1$) and, in the case of a long-period grating, a copropagating cladding mode.

Tapered fibres

Oscillatory transmissions

When the slopes of a tapered single-mode fibre are abrupt, such as those of figure B3.1, one observes oscillations in the transmitted power as a function of elongation (see figure B3.19). For a given elongation, similar oscillations occur in the transmission as a function of wavelength (see figure B3.20). This behaviour may be explained in terms of coupling and beating of the local modes of the tapered structure [12].

When the downtaper is so abrupt that the adiabaticity criterion is not fulfilled, the fundamental mode LP_{01} is unable to adapt its field and propagation constant to the guide change and it is transformed through the coupling process into a superposition of modes of the same symmetry (LP_{01}, LP_{02}, LP_{03},...), which propagate along the adiabatic central region, thus accumulating phase differences. When they enter the uptapered region, they again experience a coupling process. Power is then recovered in the core. The core transmission depends on the relative phase of the different modes. If they exit the component in phase, transmission is 1. Otherwise, power is dispatched between the fundamental core mode and cladding modes. Cladding mode power is eventually lost in the jacket.

This coupling–beatings–coupling process thus confers to a tapered fibre an oscillatory transmission. Large-amplitude oscillations seen in figure B3.20 come from LP_{01} and LP_{02} modes, while small-amplitude and larger-frequency oscillations only occur if LP_{03} and possibly higher-order modes participate in the process [13]. Besides, the LP_{01}–LP_{02} period Λ of the tapered fibre spectral

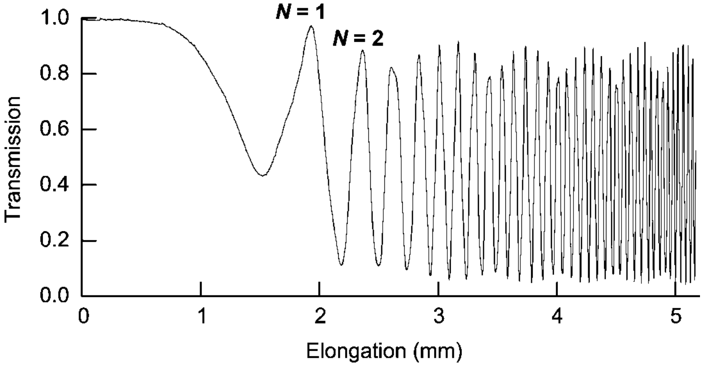


Figure B3.19. Fundamental mode transmission of a tapered single-mode fibre as a function of elongation.

response may be predicted from N, the number of oscillations observed during elongation performed at wavelength λ. For sufficiently large N, one has

$$\Lambda = \frac{\lambda}{N}. \tag{B3.20}$$

This oscillatory predictable spectral response is the basis for spectral filtering devices made of tapered fibres.

Long-period gratings (LPGs)

Gratings are also useful in terms of their filtering applications, but as opposed to tapered fibres their behaviour is based on a resonant coupling resulting in a wavelength peak in their spectral response.

- *Transfer matrix*: A sinusoidal perturbation of period $\Lambda = 2\pi/\beta_B$ induces a coupling coefficient

$$C_{12} = 2c\cos(\beta_B z + \phi) \tag{B3.21}$$

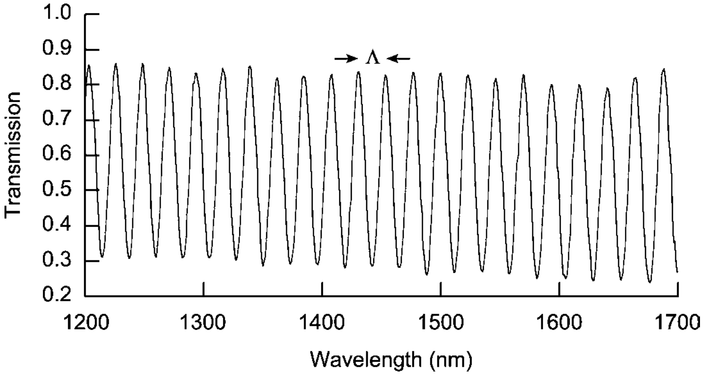


Figure B3.20. Spectral response of the tapered fibre of figure B3.19.

which, around resonance condition (B3.19), couples two codirectional modes of propagation constants β_1 and β_2. One can show, by solving coupled mode equations, that the transfer matrix is

$$\mathbf{M}_{\mathrm{LPG}} = \begin{bmatrix} \left[\cos(\gamma\ell) + \mathrm{i}\frac{\Delta}{\gamma}\sin(\gamma\ell)\right]\mathrm{e}^{\mathrm{i}(\beta_1-\Delta)\ell} & \mathrm{i}\frac{c}{\gamma}\sin(\gamma\ell)\mathrm{e}^{\mathrm{i}[(\beta_1-\Delta)\ell+\beta_{\mathrm{B}}z_0+\phi]} \\ \mathrm{i}\frac{c}{\gamma}\sin(\gamma\ell)\mathrm{e}^{\mathrm{i}[(\beta_2+\Delta)\ell-\beta_{\mathrm{B}}z_0-\phi]} & \left[\cos(\gamma\ell) - \mathrm{i}\frac{\Delta}{\gamma}\sin(\gamma\ell)\right]\mathrm{e}^{\mathrm{i}(\beta_2+\Delta)\ell} \end{bmatrix}. \tag{B3.22}$$

Here ℓ is the length of the grating extending from z_0 to $z_0+\ell$ and the other parameters are defined as follows

$$\Delta = \frac{\beta_1-\beta_2-\beta_{\mathrm{B}}}{2} = \frac{\beta_1-\beta_2}{2} - \frac{\pi}{\Lambda} \quad \text{and} \quad \gamma^2 = \Delta^2 + c^2. \tag{B3.23}$$

As for the coupler, if one assumes injection in mode 1 ($a_2(z_0)=0$), the power transfer to mode 2 is readily calculated to be

$$T_2(z_0+\ell) = |M_{21}|^2 = \frac{c^2}{\gamma^2}\sin^2(\gamma\ell). \tag{B3.24}$$

Remaining power in mode 1 can be calculated by $T_1 = |M_{11}|^2$ or using the energy conservation condition $T_1 + T_2 = 1$.

The power transfer to mode 2 is sinusoidal as a function of ℓ, but a sinc function of $\gamma = \pm\sqrt{\Delta^2+c^2}$. The peak-to-peak amplitude of the transmission is

$$T_{2\mathrm{MAX}} = \frac{c^2}{\gamma^2} = \frac{c^2}{\Delta^2+c^2}. \tag{B3.25}$$

It is always smaller than 1, being maximum for $\Delta = 0$, i.e. at resonance. Figure B3.21 shows the oscillatory fundamental mode transmission along z.

- *Bandwidth*: Let the LPG length ℓ equal half the coupling length defined by equation (B3.7): $\ell = L_{\mathrm{C}}/2 = \pi/(2c)$ so that, at resonance, the power transfer is complete. (This would occur for any odd number of $L_{\mathrm{C}}/2$.) Figure B3.22 shows, for a component of length $L_{\mathrm{C}}/2$, the power transfer from mode 1 to mode 2 as a function of the detuning $|\Delta|$.

The LPG used as a filter may be characterized by its bandwidth, i.e. by the FWHM of its transmission, or approximately of $T_{2\mathrm{MAX}}$, which in terms of Δ is equal to $2c$. It is readily converted in terms of wavelength to give

$$\delta\lambda = \frac{2\lambda^2}{L_{\mathrm{C}}|n_{\mathrm{g}1}-n_{\mathrm{g}2}|} \tag{B3.26}$$

where $n_{\mathrm{g}1}$ and $n_{\mathrm{g}2}$ are the group indices defined by

$$n_{\mathrm{g}i} = n_i - \lambda\frac{\mathrm{d}n_i}{\mathrm{d}\lambda}. \tag{B3.27}$$

The following approximation

$$\delta\lambda \approx \frac{2\lambda\Lambda}{L_{\mathrm{C}}} = \frac{\lambda}{N} \tag{B3.28}$$

is valid inasmuch as $n_{\mathrm{g}1} - n_{\mathrm{g}2} \approx n_1 - n_2$. Here, n_i are the modal effective indices ($n_i = 2\pi\beta_i/\lambda$), λ is the peak wavelength and N the total number of steps. As a result, the greater the number of steps, the more

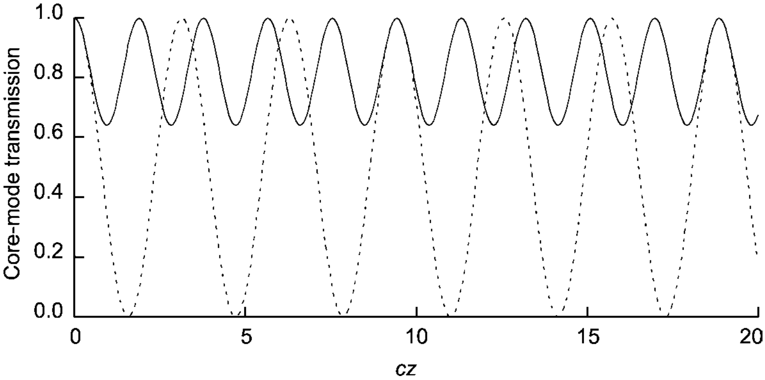


Figure B3.21. Fundamental mode transmission along the grating for different values of the detuning parameter Δ. The oscillation amplitude is maximum (equal to 1) for $\Delta = 0$ and the frequency minimum (dotted line). As Δ increases, γ increases as well, making the amplitude decrease and the frequency increase: the transmission is shown for $\Delta/c = 4/3$ (plain line).

selective the grating will be. For a given grating step (determined by the pair of chosen modes at λ), the longer the grating (and as a consequence the smaller the coupling coefficient c, since $L = L_C/2$), the narrower its bandwidth.

- *Fundamental mode transmission*: The cladding modes being absorbed by the jacket, most of the time, one only has access to T_1. LPGs are thus rejection band filters. Actually, for a given grating step Λ, several pairs of modes (i,j) may be resonant in a wavelength range inasmuch as the following condition is fulfilled.

$$\beta_i - \beta_j = \frac{2\pi}{z_{ij}} = \frac{2\pi}{\lambda}(n_i - n_j) = \frac{2\pi}{\Lambda}. \tag{B3.29}$$

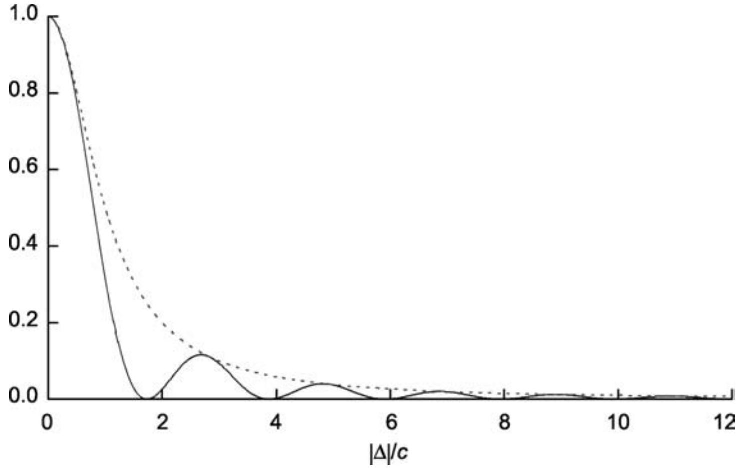


Figure B3.22. Power transfer T_2 as a function of the detuning $|\Delta|$ in a component of length $L_C/2$ (plain line). Its amplitude $T_{2\mathrm{MAX}}$ is the dotted line.

Here, z_{ij} is the beating length of modes i and j and n_i and n_j their effective indices, which depend on the wavelength λ. This equation generalizes the matching condition written in equation (B3.19). If, as in tapered fibres, the perturbation preserves the circular symmetry, only modes of the same symmetry are involved. Figure B3.23 shows the beating lengths z_{1j} for $LP_{01}-LP_{0j}$ modal pairs as functions of λ. This graph permits us to find the resonance wavelengths for a given step Λ. Alternatively, it gives the periods that may be chosen to realize a filter at a given wavelength.

Figure B3.24 gives, as a function of wavelength, the dB transmission in the fundamental mode of a grating of step $\Lambda = 525\ \mu m$ in a standard telecommunication fibre.

Note finally that, with slanted gratings, it is also possible to transfer power from the fundamental core mode LP_{01} to cladding modes with different symmetry such as LP_{11} and LP_{21}.

The main limitation of the LPGs used as filtering components is their temperature dependence of the order of $100-150\ \text{pm}\ {}^\circ\text{C}^{-1}$ (10–15 nm/100°C). This high figure, which largely depends on the fibre, comes from the fact that the principle of operation of an LPG is based on the coupling of a core mode with a cladding mode. Being made up of different materials, these two layers are subjected during hot drawing and subsequent cooling of fibre to different stresses, which result in a different thermal sensitivity. LPGs made of standard SMF28 *Corning* fibres exhibit a thermal dependence of $50\ \text{pm}\ {}^\circ\text{C}^{-1}$.

Short-period gratings

Let the period $\Lambda = 2\pi/\beta_B$ of the sinusoidal perturbation couple two contradirectional modes of the same propagation constant $\beta_1 = -\beta_2 = \beta$. The corresponding coupling coefficient may be written as

$$C_{12} = 2ic\cos(\beta_B z + \phi). \tag{B3.30}$$

With a notation similar to that used in the codirectional case, one has

$$\Delta = \frac{\beta_1 - \beta_2}{2} - \frac{\beta_B}{2} = \beta - \frac{\beta_B}{2} = \beta - \frac{\pi}{\Lambda}.$$

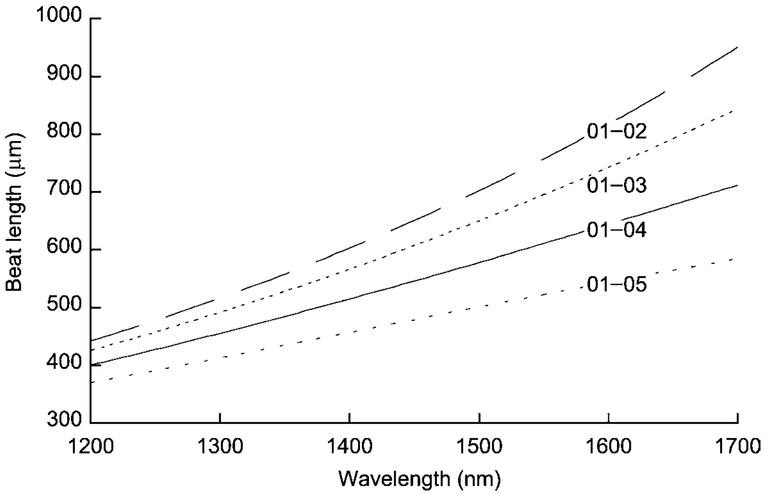


Figure B3.23. Beating lengths of $LP_{01}-LP_{0j}$ modal pairs for a typical telecommunication fibre.

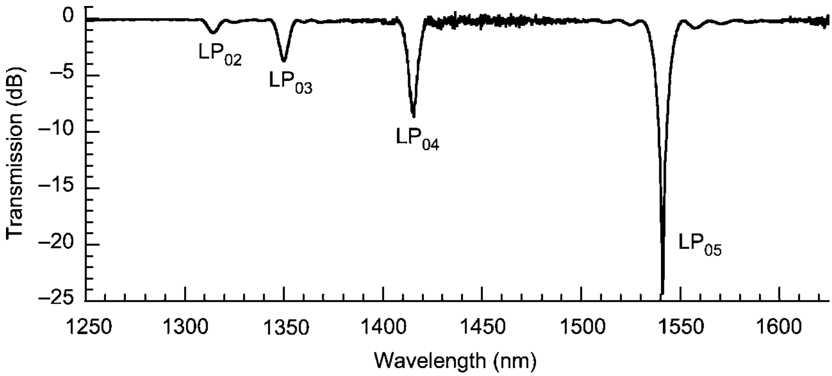


Figure B3.24. Fundamental mode transmission of a long-period grating. Experimental response (plain line) is compared to the calculated one (dotted line). The grating was made by irradiating a standard telecommunication fibre with a CO_2 laser. It consists of 118 steps of 525 μm. The peaks are identified by the $0m$ indices of the LP cladding mode, which is coupled with the fundamental core mode.

- *Transfer matrix*: By solving the coupled mode equations, one then finds the transfer matrix of a grating extending from z_0 to $z_0 + \ell$

$$\mathbf{M}_{\mathrm{SPG}} = \begin{bmatrix} \left[\cosh(\gamma\ell) + \mathrm{i}\frac{\Delta}{\gamma}\sinh(\gamma\ell)\right]\mathrm{e}^{\mathrm{i}(\beta-\Delta)\ell} & \mathrm{i}\frac{c}{\gamma}\sinh(\gamma\ell)\mathrm{e}^{\mathrm{i}[(\beta-\Delta)\ell+\beta_{\mathrm{B}}z_0+\phi]} \\ -\mathrm{i}\frac{c}{\gamma}\sinh(\gamma\ell)\mathrm{e}^{-\mathrm{i}[(\beta-\Delta)\ell+\beta_{\mathrm{B}}z_0-\phi]} & \left[\cosh(\gamma\ell) - \mathrm{i}\frac{\Delta}{\gamma}\sinh(\gamma\ell)\right]\mathrm{e}^{-\mathrm{i}(\beta-\Delta)\ell} \end{bmatrix} \tag{B3.31}$$

where

$$\gamma^2 = c^2 - \Delta^2. \tag{B3.32}$$

1. If $\Delta \leq c$,

$$\gamma = \sqrt{c^2 - \Delta^2} \tag{B3.33}$$

and solutions are those written in equation (B3.31) with γ real.

2. If $\Delta \geq c$,

$$\gamma = \mathrm{i}\sqrt{\Delta^2 - c^2} \tag{B3.34}$$

so that solutions can be written as in equation (B3.31), but with $\gamma = \mathrm{i}|\gamma|$ an imaginary number. Using then $\cosh(\mathrm{i}|\gamma|z_0) = \cos(|\gamma|z_0)$ and $\sinh(\mathrm{i}|\gamma|z_0) = \mathrm{i}\sin(|\gamma|z_0)$, one can rewrite them as sinusoidal functions.

- *Fundamental mode reflection and transmission*: The reflection and transmission coefficients may then be calculated for given limit conditions. Inversion of the matrix $\mathbf{M}_{\mathrm{SPG}}$ gives

$$\begin{bmatrix} a_1(z_0) \\ a_2(z_0) \end{bmatrix} = \begin{bmatrix} M_{22} & -M_{12} \\ -M_{21} & M_{11} \end{bmatrix} \begin{bmatrix} a_1(z_0+\ell) \\ a_2(z_0+\ell) \end{bmatrix}. \tag{B3.35}$$

Let $z_0 = 0$. Usually one has $a_2(\ell) = 0$, so that one finds the amplitude reflection and transmission coefficients

$$r = \frac{a_2(0)}{a_1(0)} = -\frac{M_{21}}{M_{22}} \quad \text{and} \quad t = \frac{a_1(\ell)}{a_1(0)} = \frac{1}{M_{22}} \tag{B3.36}$$

and the power coefficients

$$R = R(0) = \frac{|M_{21}|^2}{|M_{22}|^2} = \frac{|M_{21}|^2}{|M_{11}|^2} \quad \text{and} \quad T = T(\ell) = \frac{1}{|M_{22}|^2} = \frac{1}{|M_{11}|^2} \tag{B3.37}$$

Explicitly, one has to distinguish the cases depending on the detuning with respect to the coupling coefficient.

1. If $\Delta \leq c$

$$R(z) = \frac{|a_2(z)|^2}{|a_1(0)|^2} = \frac{c^2}{\gamma^2} \frac{\sinh^2[\gamma(\ell - z)]}{\cosh^2(\gamma\ell) + \frac{\Delta^2}{\gamma^2}\sinh^2(\gamma\ell)} \tag{B3.38}$$

$$T(z) = \frac{|a_1(z)|^2}{|a_1(0)|^2} = \frac{\cosh^2[\gamma(\ell - z)] + \frac{\Delta^2}{\gamma^2}\sinh^2[\gamma(\ell - z)]}{\cosh^2(\gamma\ell) + \frac{\Delta^2}{\gamma^2}\sinh^2(\gamma\ell)} \tag{B3.39}$$

with $\gamma = \sqrt{c^2 - \Delta^2}$.

2. If $\Delta \geq c$

$$R(z) = \frac{|a_2(z)|^2}{|a_1(0)|^2} = \frac{c^2}{\gamma^2} \frac{\sin^2[\gamma(\ell - z)]}{\cos^2(\gamma\ell) + \frac{\Delta^2}{\gamma^2}\sin^2(\gamma\ell)} \tag{B3.40}$$

$$T(z) = \frac{|a_1(z)|^2}{|a_1(0)|^2} = \frac{\cos^2[\gamma(\ell - z)] + \frac{\Delta^2}{\gamma^2}\sin^2[\gamma(\ell - z)]}{\cos^2(\gamma\ell) + \frac{\Delta^2}{\gamma^2}\sin^2(\gamma\ell)} \tag{B3.41}$$

with $\gamma = \sqrt{\Delta^2 - c^2}$.

The minimum transmission (corresponding to a maximum reflection) takes place at resonance: $\Delta = 0$ or $\gamma = c$ (figure B3.25). For a grating of length ℓ one calculates

$$R_{\text{Max}} = \tanh^2(c\ell) \quad \text{and} \quad T_{\min} = \frac{1}{\cosh^2(c\ell)} \tag{B3.42}$$

Figures B3.26 and B3.27 show the reflection and transmission behaviour as a function of z depending on the value of the detuning parameter Δ. The transmission coefficient $T(z)$ may be locally >1. This local energy accumulation, similar to that observed in a Fabry–Perot interferometer, is due to the resonance or stationary waves between the grating ends. This effect, which gives rise to side lobes in the spectral response, is undesirable for spectral filtering applications. It may be suppressed by apodizing techniques as explained in a subsequent section.

- *Spectral response*: As for LPGs, the wavelength response is determined by the transmission and reflection coefficients as functions of the Δ parameter. The magnitude of Δ compared to that of the coupling coefficient c determines γ as per equation (B3.32), which is real close to the resonance condition, but purely imaginary as soon as $\Delta > c$. Transmission then becomes oscillatory, resulting in the typical spectral side lobes. The $\Delta = c$ condition, as in the LPG case, may thus be used to define the peak width. An example spectral response in reflection is shown in figure B3.28.

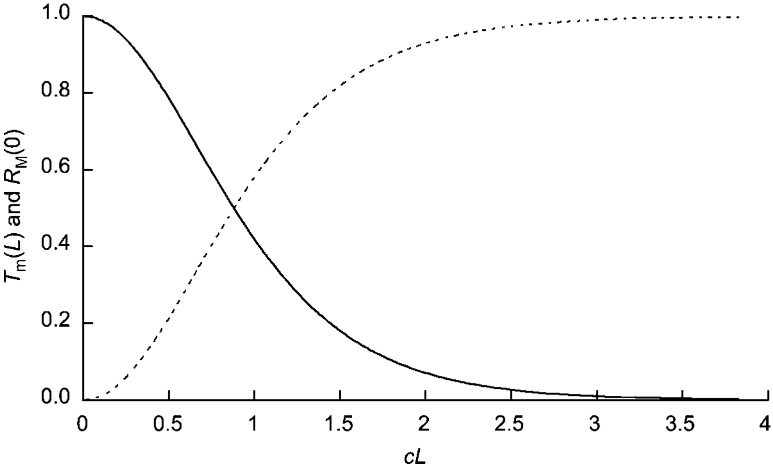


Figure B3.25. Reflection and transmission in the fundamental core mode at resonance ($\Delta = 0$) as a function of the grating length.

- *Apodization*: For most applications, side lobes seen in the spectral response are undesirable. They originate from the multiple reflections that take place between the grating ends as in a Fabry–Perot resonator. One can also understand this effect by remembering that, in a first approximation (i.e. when the perturbation is weak), the spectral response of a grating is the Fourier transform of the amplitude c of the refractive index modulation. The Fourier transform of a rectangular function is a sinc function. The larger the rectangular function, the narrower the sinc function. Similar observations are valid for the grating: the longer the grating, the narrower the spectral width and its side lobes.

One can reduce these side lobes by smoothing the coupling coefficient, giving it a gaussian envelope. This results in smoothing the wavelength response (in other words, apodizing). To be efficient both on the short- and long-wavelength sides, the average refraction index all over the grating length should be uniform. Figure B3.29 shows the effect of apodization on the grating of figure B3.28.

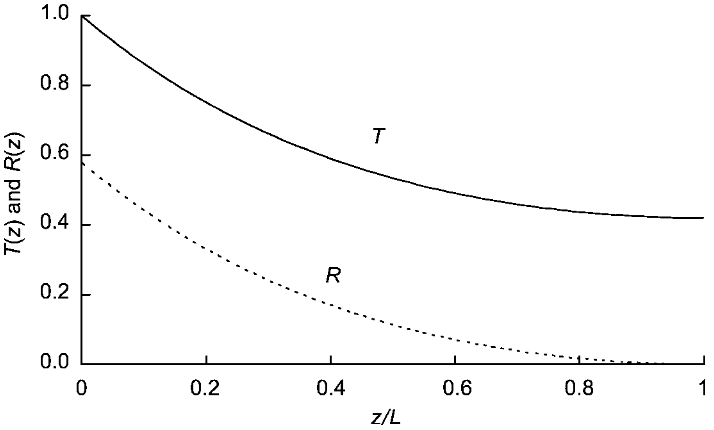


Figure B3.26. Reflection $R(z)$ and transmission $T(z)$ in the fundamental mode along the grating of length $\ell = 1/c$ at resonance, i.e. for $\Delta = 0$. Since $\Delta < c$, the amplitude varies exponentially.

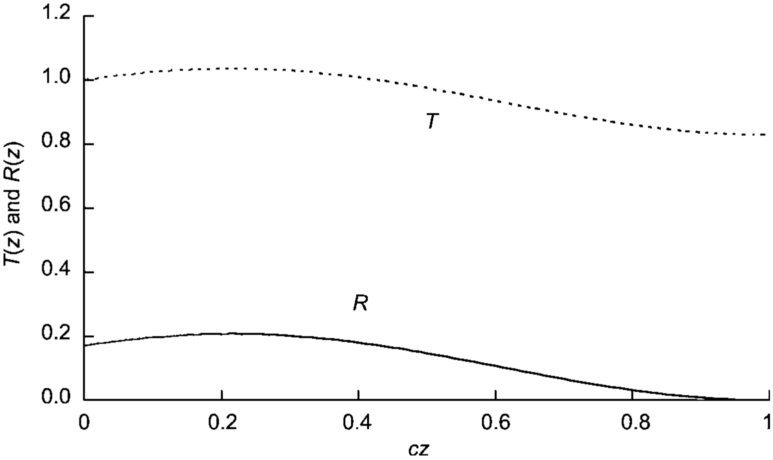


Figure B3.27. Reflection $R(z)$ and transmission $T(z)$ in the fundamental mode along the grating for a far from resonance case $\Delta > c$. In this case, the amplitude shows a sinusoidal variation. The calculation is made for $\Delta/c = \sqrt{17}$.

The thermal dependence of SPGs is much weaker than that of LPGs. It is typically 10 pm °C^{-1} (1 nm/100°C), i.e. approximately ten times less than that of LPGs, and does not depend on the fibre. Nevertheless, the thermal sensitivity of these gratings, irrespective of their application, must be compensated by an appropriate packaging.

Spectral filters

Spectral filtering is a key function of WDM networks. Some applications are described below.

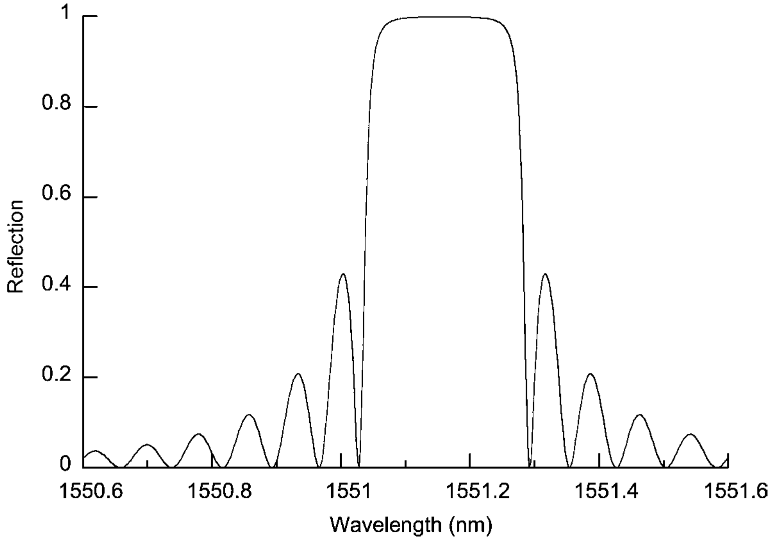


Figure B3.28. Power reflection coefficient of a short-period grating as a function of wavelength (simulation).

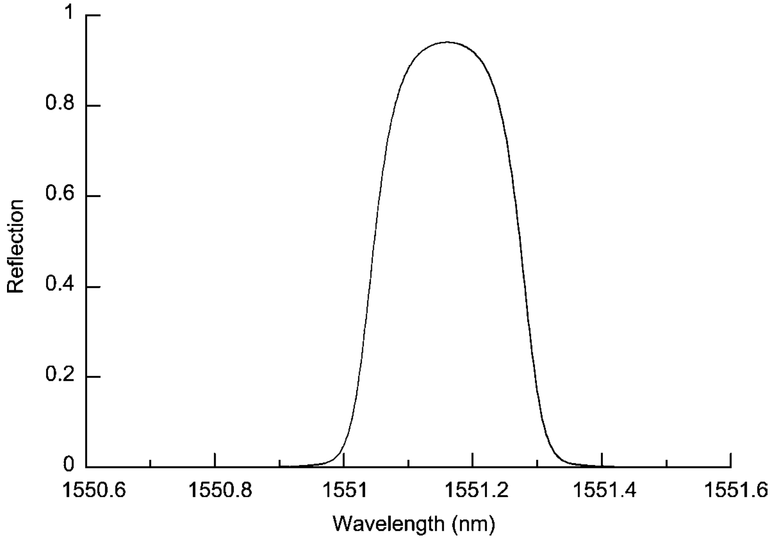


Figure B3.29. Power reflection coefficient of an apodized short-period grating as a function of wavelength (simulation).

Narrow-width spectral filters

The most immediate application of the Bragg SPGs is to use them as selective mirrors in all-fibre lasers in linear (as opposed to loop) configuration. They are currently found in doped fibre lasers as well as cascaded Raman lasers used to pump Raman gain amplifiers.

Bragg SPGs are also used to stabilize both the wavelength and power of semiconductor pump laser diodes. In this case, the grating, with a reflection coefficient of only a few per cent, is integrated in the pigtail at a distance exceeding the coherence length of the laser.

Finally, filters are needed to enhance isolation between the pump and the signal or between two adjacent signal channels. They are also used in combination with a circulator (which is a micro-optic device) to drop the signal of a particular wavelength channel. The ideal spectral amplitude shape to accomplish this type of function is then a rectangular response with a prescribed bandwidth. The Bragg reflection grating remains well suited for this type of filtering, especially if narrowness is an issue: although not perfectly rectangular, well-known standard apodization techniques are used to eliminate undesirable side lobes and give rise to excellent spectral characteristics. However, as can be seen in figure B3.30 (top), such a standard grating filter has a parabolic group delay profile that adversely affects signals at the transmission rates of 10 and 40 Gb s^{-1}. More complex apodization profiles with phase shifts allow for correction and equalization of this parabolic group delay characteristic and result in an ultra-low dispersion over the filter passband. As shown in figure B3.30 (bottom), group delay can be minimized to a ripple function of less than ± 5 ps in amplitude [14].

Large-width spectral filters

Other filters with larger bandwidths are also needed. Among them, the most often used is the gain flattening filter (GFF) designed to equalize the channels amplified by Er-doped fibre gain. The following all-fibre technologies have been explored and implemented sometimes in conjunction.

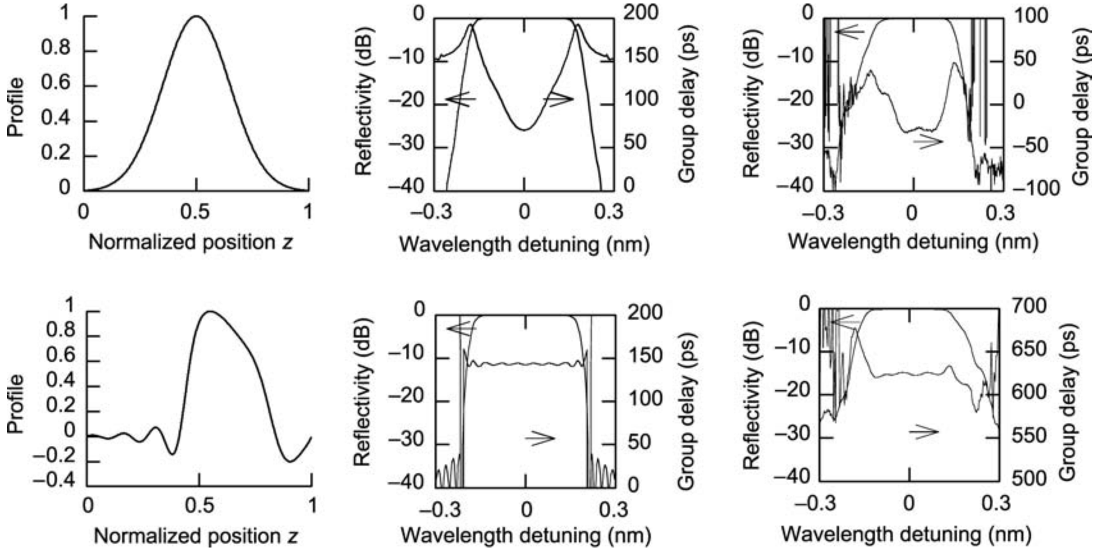


Figure B3.30. Comparison of standard (top) and low-dispersion (bottom) designs of Bragg SPG 50 GHz filters. The apodization profile (left column), theoretical reflectivity and group delay spectra (centre column) and experimental measurements (right) are compared (by courtesy of TeraXion).

1. fusion-tapering technology;
2. short-period gratings (SPGs), straight or slanted;
3. long-period gratings (LPGs), straight or slanted;
4. thin-film technology.

However, the best fits to the inverse gain curve are obtained by using concatenated tapered structures or chirped short-period unslanted Bragg gratings in the transmission mode. Both types of GFF offer similar performance in terms of error function ($<\pm 0.1$ dB), PMD (<0.05 ps). Thermal stability is better for gratings ($<0.5\,\mathrm{pm\,°C^{-1}}$) than for tapered fibre filters ($<2\,\mathrm{pm\,°C^{-1}}$). While the grating solution is more compact, it suffers back reflection, which needs an additional isolator. Group delay ripples, which result from undesired weak reflections occurring along the grating, are negligible with a typical amplitude of ± 0.3 ps. They both are commercially available. Example responses are shown in figures B3.31 and B3.32.

Mode converters

At the resonance wavelength, an LPG is a mode converter operating in the codirectional scheme. Similarly, an SPG can also be used as a mode converter, but operates in the contradirectional scheme, since not only the fundamental mode, but other contradirectional modes may also be reflected through a short-period Bragg grating [15]. By choosing the step, the length and the grating tilt with respect to the fibre axis, one can choose the wavelength and the mode(s) in which there is conversion. Figure B3.33 shows the spectral response of an LP_{01}–LP_{11} converter. To avoid undesirable conversion to cladding modes,

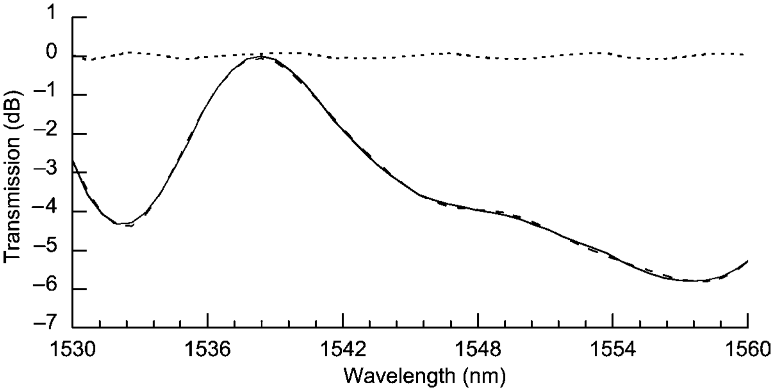


Figure B3.31. Transmission of a GFF made with the fusion-tapering technique. The plain line shows the target filter transmission, while the dashed one is that of the manufactured filter. The dotted line around 0 dB shows the error function. Its amplitude is less than ±0.1 dB (by courtesy of ITF Optical Technologies).

a special fibre should be used. Used in conjunction with mode splitters, the mode converters open up the design of new all-fibre components, which would get rid of micro-optic devices such as circulators.

OADMs

A signal at the Bragg wavelength rather than reflected from a short-period grating back to the source can be extracted. It is then a selective wavelength demultiplexer, which has a significant role to play in future communication networks. Indeed, the strategy that is implemented to currently increase the data flow in a fibre link is the DWDM. Channel frequency spacing is typically 100 GHz (corresponding to 0.8 nm around $\lambda = 1550$ nm), which makes it possible to put 40 channels in the EDFA gain bandwidth

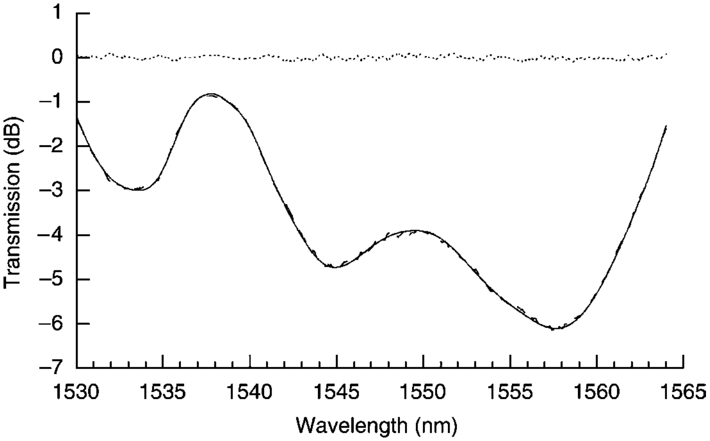


Figure B3.32. Transmission of a GFF made of a chirped Bragg grating. The plain line shows the target filter transmission, while the dashed one is that of the manufactured filter. The dotted line around 0 dB shows the error function. Its amplitude is less than ±0.1 dB (by courtesy of TeraXion).

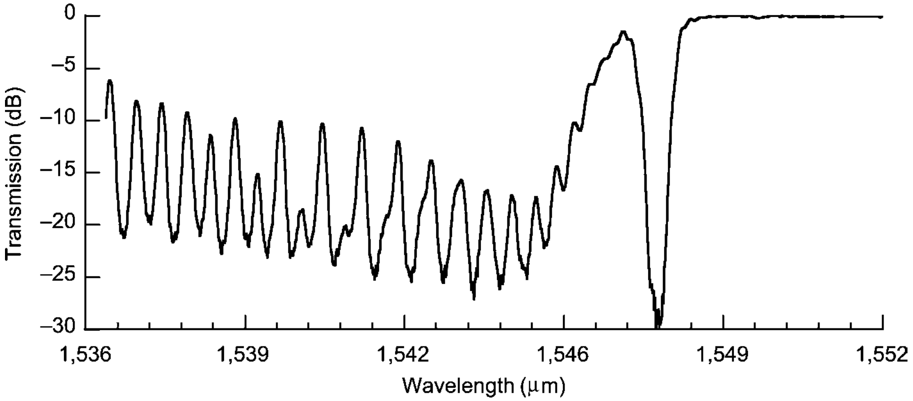


Figure B3.33. Experimental transmission of a mode converter as a function of wavelength. The component consists in a slanted short-period grating written in a bimodal fibre. The fundamental LP_{01} core mode is converted into the LP_{11} core mode at $\lambda = 1547.5$ nm. At shorter wavelengths conversion to higher-order LP_{1m}, LP_{2m}, LP_{3m} cladding modes can be seen.

(the C-band) and thus to multiply the current flow by 40, which is of 2.5 Gbit s^{-1}. Narrower spacings such as 50 and 25 GHz (0.4 and 0.2 nm) are also used (see figure B3.16 for an example 50 GHz spacing component).

To extract the reflected signal, one can use a circulator, which is made up of micro-optic components. An alternative all-fibre solution is shown in figure B3.34, which makes use of identical gratings in an MZ arrangement. Cross-talk between channels must be avoided by apodizing the spectral response and by a proper design of the fibre to get rid of the cladding mode reflections. From the fabrication viewpoint, the symmetry of the component is critical for its correct operation. Alternative solutions making use of a single grating written in the central region of a single coupler were also studied [16].

Dispersion compensators

Dispersion in standard fibres mainly comes from material CD, which happens to be zero around $\lambda = 1.3\ \mu m$. In the C-band (around $\lambda = 1.55\ \mu m$), dispersion causes a pulse to spread in time, the highest frequencies (shortest wavelengths) arriving before the lowest ones. CD is inherently troublesome and solutions to eliminate it completely by using dispersion-shifted fibre were rapidly abandoned because

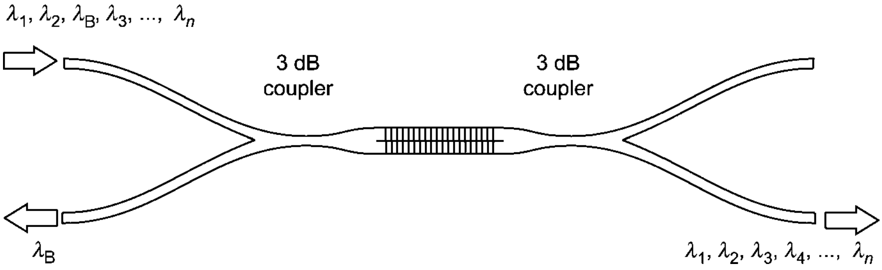

$\lambda_1, \lambda_2, \lambda_B, \lambda_3, \ldots, \lambda_n$
3 dB coupler
3 dB coupler
λ_B
$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_n$

Figure B3.34. Principle of operation of an all-fibre OADM.

four-wave mixing (a non-linear effect, which induces cross-talk between channels) is automatically phase matched around the zero-dispersion wavelength. It is thus better to compensate for dispersion by using additional devices.

The simplest way is to use a dispersion compensation module (DCM), which consists of a length of a different fibre with a carefully designed index profile to tailor the modal propagation constant. A relatively long fibre length of compensating fibre must be used (typically 20 km of compensating fibre for 100 km of standard fibre), and it only partially compensates for higher-order dispersion effects (typically 60% of the slope), which is the channel-to-channel variation of the dispersion.

Because of their compactness, Bragg SPGs are alternative promising devices that may be used for dispersion compensation. Chirped SPGs (CSPGs) shown in figure B3.35 reflect different frequency components of a pulse at different locations. Its first step being longer than the last, with a careful design, it can compensate for the delay of the low frequencies on the high ones. To recover the reflected signal, a circulator (a micro-optic component) is necessary. An SPG can have a dispersion of several orders of magnitude higher than a similar length of fibre, so that a few cm long dispersion compensator compensates for the dispersion of many kilometres of optical fibre. While relatively narrowband in nature, CSPGs can be made wideband by making them multi-channel. A multi-channel CFBG-based dispersion compensator can be made by writing many CSPG components in separate sections of a fibre or by superimposing many CSPG components on the same section of fibre, the latter solution resulting in a very compact component. Another advantage of the SPGs over standard broadband compensation techniques such as DCF is that they are inherently independent of each other, which allows the design of complex structures including channel skipping and channel-to-channel variation of the dispersion.

Using the superimposed approach, compensation over up to 32 channels has been achieved with 100 GHz spacing and allowing for compensation of both the second- and third-order CD accumulated

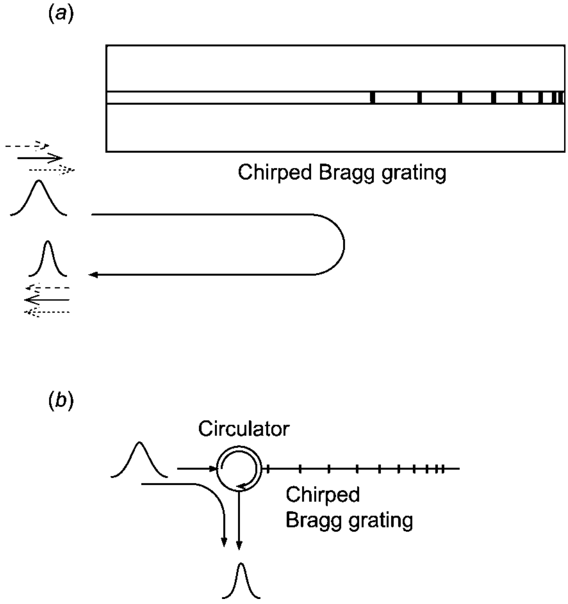


Figure B3.35. (*a*) Principle of operation of a chirped SPG to compensate pulse dispersion; (*b*) dispersion compensator made of a chirped SPG associated with a circulator.

over 50 km of SMF-28 fibre [14]. Due to the possible adjustment of the dispersion on a per-channel basis, the wideband compensation based on a superimposed grating is a good candidate for compensating such a residual CD. As an example, figure B3.36 shows the reflectivity and group delay spectra of a 32-channel dispersion compensation grating with a 100 GHz spacing in which the dispersion varies from channel to channel: from -860 to $-800\,\mathrm{ps\,nm^{-1}}$.

CSPGs can also be used in conjunction with DCM to fully manage the dispersion over a large number of WDM channels.

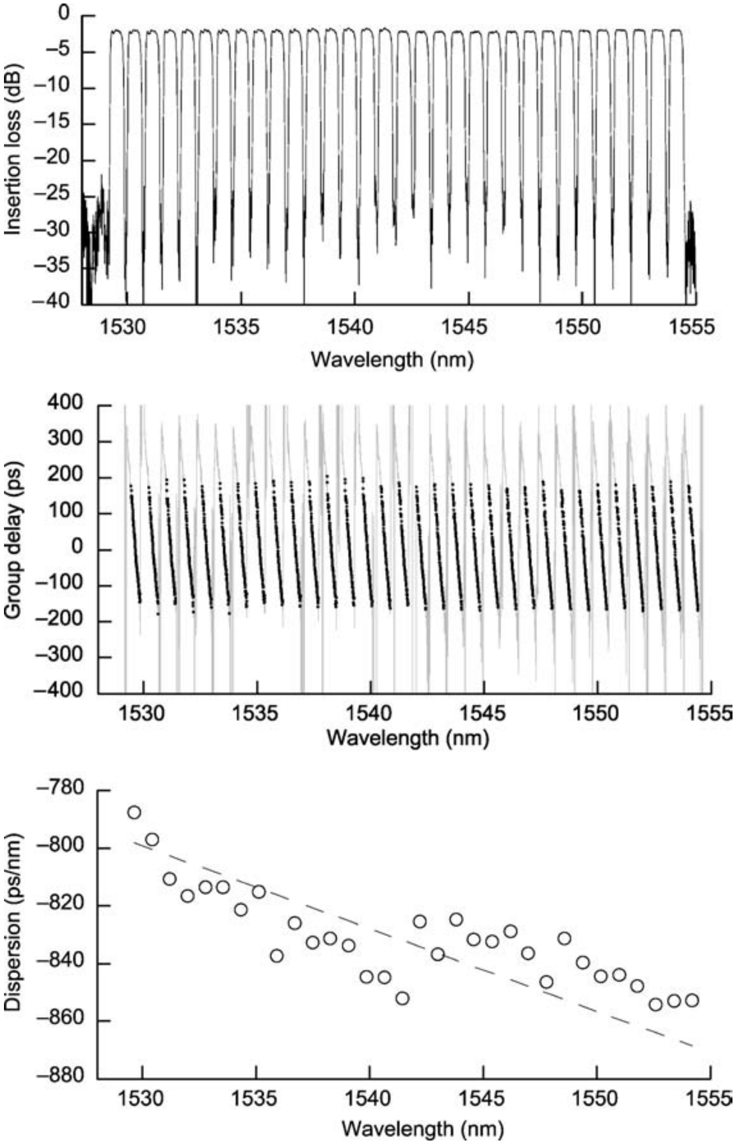


Figure B3.36. Reflectivity spectrum (top), group delay spectrum (centre), and dispersion (bottom) of a 32-channel dispersion compensation grating (by courtesy of TeraXion).

B3.5 Conclusion

This chapter has highlighted the importance of all-fibre components, which are an integral part of the current telecommunication networks. They already perform extremely diversified functions and their performance, thanks to the maturity of manufacturing technologies, is excellent. Depending on applications, they must sometimes compete with their integrated or micro-optics equivalents. The future communication networks will undoubtedly take advantage of all the available technologies, but the all-fibre will keep a privileged position thanks to the inherent low excess loss.

Acknowledgments

The authors gratefully acknowledge the contribution of Dr Nawfel Azami and Dr Nicolas Godbout from ITF Optical Technologies Inc. and of Dr Yves Painchaud from TeraXion.

References

[1] Stewart W J and Love J D 1985 Wavelength-flattened fused couplers *Electron. Lett.* **21** 742–743

[2] Lacroix S, Gonthier F and Bures J 1988 Fibres unimodales effilées *Ann. Télécomm.* **43** 43–47

[3] Kashyap R 1999 Fiber Bragg gratings *Optics and Photonics* ed T Tamir (New York: Academic)

[4] Othonos A and Kalli K 1999 *Fiber Bragg Gratings, Fundamentals and Applications in Telecommunication and Sensing* (Boston, MA: Artech)

[5] Perron D, Orsini P, Daxhelet X, Lacroix S and Verhaegen M 2000 Long-period-grating fabrication techniques *ICAPT 2000—Photonics North*

[6] Snyder A W and Love J D 1983 *Optical Waveguide Theory* (London: Chapman and Hall)

[7] Lacroix S, Gonthier F and Bures J 1994 Modeling of symmetric 2 × 2 fused-fiber couplers *Appl. Opt.* **33** 8361–8369

[8] Birks T A and Hussey C D 1989 Wavelength-flattened couplers: performance optimisation by twist-tuning *Electron. Lett.* **25** 407–408

[9] Gonthier F, Ricard D, Lacroix S and Bures J 1991 Wavelength-flattened 2 × 2 splitters made of identical single-mode fibers *Opt. Lett.* **16** 1201–1203

[10] Azami N, Villeneuve E, Villeneuve A and Gonthier F 2003 All-sop all-fiber depolarizer linear design *Optical Fiber Conference TuK5* pp 230–231

[11] Shou Y, Bures J, Lacroix S and Daxhelet X 1999 Mode separation in fused fiber coupler made of two-mode fibers *Opt. Fiber Technol.* **5** 92–104

[12] Love J D, Stewart W J, Henry W M, Black R J, Lacroix S and Gonthier F 1991 Tapered single-mode fibres and devices: Part 1. Adiabaticity criteria *EE Proc. Pt. J: Optoelectron.* **18** 343–354

[13] Lacroix S, Bourbonnais R, Gonthier F and Bures J 1986 Tapered monomode optical fibers: understanding large power transfer *Appl. Opt.* **25** 4424–4429

[14] Lachance RL, Painchaud Y and Doyle A 2002 Fiber Bragg gratings and chromatic dispersion *ICAPT 2002—Photonics North* pp 1009–1016

[15] Vaillancourt I, Daxhelet X, Lacroix S and Godbout N 2001 A 99.9% efficient lp_{01}–lp_{11} mode converter without lp_{01} back reflection: a detailed analysis *Bragg Gratings, Photosensitivity, and Poling in Glass Waveguides, OSA* pp PD4–1

[16] Bakhti F, Sansonetti P, Sinet C, Gasca L, Martineau L, Lacroix S, Daxhelet X and Gonthier F 1997 Novel optical add/drop multiplexer based on uv-written Bragg grating in a fused 100% coupler *Electron. Lett.* **33** 803–804

B4
Optical modulators

Nadir Dagli

B4.1 Introduction

Optical modulators accept as input the CW or pulsed output of the laser diode and generate as output a modulated optical waveform. The modulating signal is an electrical input voltage either in digital or analogue format. Using an external modulator allows the laser to operate independently. Hence its output power and frequency is controlled very accurately. There are many different applications that require modulators. One very important application is the digital fibre optic communication systems. The advent and widespread use of the Internet increased the data transmission rates drastically. The demand for more bandwidth continues on an ongoing basis. Fibre optic communication networks have the capability to deliver such demand. At present, 40 Gbit s^{-1} fibre optic transmission systems are being developed and installed all around the globe. This requires the modulation of the optical signal at such rates. Since most data generated are in electrical form, some form of electrical to optical modulation is required. This can be achieved by directly modulating the output of a laser diode through its drive current. There are laser diodes with small signal modulation bandwidths approaching 40 GHz [1]. Although this approach is simple, typically it is not used over 2.5 Gbit s^{-1} mainly because of the chirp of the directly modulated laser output [5]. The undesired frequency modulation associated with amplitude modulation severely limits the transmission distance over the fibre at high bit rates. This difficulty necessitated the development of external optical modulators. Transmission of analogue signals also requires external modulators. Analogue transmission is typically used to carry and distribute analogue cable TV signals. Such distribution eliminates frequent microwave amplification needed on a coaxial distribution system, improving the reach and reliability [2]. Subcarrier multiplexing is another scheme that allows transmission independent of data format. It can be combined with digital transmission to improve functionality, such as optical labelling and header recognition [3]. Modulators are also used in military applications, especially in phased array radar [4]. Control signals to different radiating elements can be carried over the optical fibre, allowing for large separation between the antenna and the control site in radar. Furthermore, much other functionality in the microwave domain, such as tunable filtering, tunable delaying and high-speed analogue to digital conversion, can be performed using photonics techniques. All these microwave photonics applications rely on optical modulators for electrical to optical conversion. However, the required properties of the modulator for each one of these applications are different. Digital applications typically require low drive voltages, wide bandwidth and adjustable chirp. On the other hand, analogue applications require a high degree of linearity and very low insertion loss. These requirements are most often conflicting and impose significant challenges on the modulator design and fabrication.

This chapter starts with the description of the criteria used to characterize an optical modulator. Next fundamentals of phase and amplitude modulators are given. After that the travelling wave modulation technique, which is universally used in all high-speed modulators, is described. This is followed by the description of physical effects used in optical modulation. Both electroabsorptive and electro-optic effects are described. Then specific examples of modulators in different material systems are given. These are electroabsorption, $LiNbO_3$, III–V compound semiconductor, and polymer modulators. Finally, a brief summary is presented.

B4.2 Modulator specifications

A modulator, which is described as a block diagram shown in figure B4.1, is characterized using certain specifications. These are defined and described briefly in the following subsections. Their ranges and specifics for different technologies are described in detail later on.

B4.2.1 Insertion loss

This is the optical power loss in the on state. It is typically given in units of decibels and is defined as

$$10\log\left(\frac{(P_{\mathrm{Out}})_{\mathrm{On}}}{P_{\mathrm{In}}}\right) \tag{B4.1}$$

where $(P_{\mathrm{Out}})_{\mathrm{On}}$ and P_{In} are the output power in the on state and input power, respectively.

B4.2.2 Extinction ratio

This is the ratio of the off state output power, $(P_{\mathrm{Out}})_{\mathrm{Off}}$, to the on state output power and typically is defined in decibel units as

$$10\log\left(\frac{(P_{\mathrm{Out}})_{\mathrm{Off}}}{(P_{\mathrm{Out}})_{\mathrm{On}}}\right). \tag{B4.2}$$

B4.2.3 Drive voltage

This is the voltage required to switch the modulator from on to off state. This voltage is most often referred to as V_π. V_π is desired to be as low as possible, especially for high-speed operation, where the generation of voltages larger than a few volts may be very difficult. Most commercial modulators require a modulator driver, which amplifies the voltage available to a level sufficient to drive the modulator at the required impedance level.

B4.2.4 Chirp

In a modulator, amplitude and phase changes are coupled. That means that every time amplitude changes so does the phase and vice versa. This coupling could be due to material properties and/or

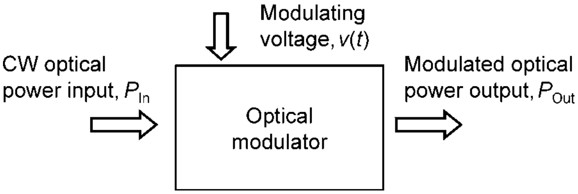

Modulating voltage, v(t)
CW optical power input, P_In
Optical modulator
Modulated optical power output, P_Out

Figure B4.1. Block diagram of an optical modulator.

modulator geometry. Since frequency is time rate of change of phase, phase modulation accompanying amplitude modulation changes the instantaneous frequency of the optical wave. This is known as chirping. Chirping combined with dispersion could severely limit the transmission distance of high-speed pulses in a communication system [5]. The chirp of a modulator is quantified using a chirp parameter, σ, which is the ratio of phase modulation to amplitude modulation, as

$$\sigma = \frac{\dfrac{\mathrm{d}\phi}{\mathrm{d}t}}{\dfrac{1}{|\mathscr{E}|}\dfrac{\mathrm{d}|\mathscr{E}|}{\mathrm{d}t}} \tag{B4.3}$$

where ϕ and $|\mathscr{E}|$ are the phase and the amplitude of the optical electric field at the output of the modulator. The chirp parameter is expected to be zero for a pure amplitude modulator and infinity for a pure phase modulator.

B4.2.5 Polarization dependence

This quantifies the performance of the modulator for different polarizations. Since any incoming polarization can be considered as the superposition of two mutually orthogonal polarizations, usually the performance of the modulator is specified with respect to two mutually orthogonal polarizations. These polarizations are chosen as transverse electric (TE) and transverse magnetic (TM) in a guided wave modulator. The physical effect used in most modulators is observed only for certain polarizations of the optical mode. In other words, the operation of the modulator is polarization dependent. This is not a major concern since the modulator is used right after the laser diode. However, this could become an important issue if the modulator is used after some fibre transmission, which can generate random polarization at the input of the modulator. Such difficulty can be dealt with using polarization diversity schemes, which separates the random output polarization into two mutually orthogonal polarizations and deal with each polarization component separately.

B4.2.6 Bandwidth

Bandwidth specifies the range of modulation frequencies over which the device can be operated. The bandwidth is usually taken as the difference between the upper and lower frequencies at which the modulation depth falls to 50% of its maximum value.

B4.2.7 Bias stability and drift

For some modulators the bias voltage required to keep the modulator properly operating may change over time. Therefore, constant monitoring of the bias point and its continuous adjustment using a feedback circuitry may be required.

B4.3 Phase modulators

As described in section B4.6, the index of refraction of a material can be changed using external perturbations. Changes in the index of refraction create changes in the phase velocity and the phase of the optical wave. Therefore, an optical wave propagating in a material whose index of refraction is modulated is phase modulated. Phase modulation can be achieved either in bulk material or in an optical waveguide.

In either case, any component of the electric field of the propagating optical wave can be expressed as

$$\mathscr{E}(z,t) = \mathscr{E}_0 \cos(\omega t - \beta z) = \mathscr{E}_0 \cos(\omega t - \frac{2\pi}{\lambda_0} nz). \tag{B4.4}$$

If the index of refraction n is perturbed using a time varying external voltage $v(t)$

$$n(t) = n_0 + \Delta n = n_0 + Kv(t) \tag{B4.5}$$

where K is a proportionality constant depending on the physical effect, geometry and the material used. Specific K values depend on the technology and are described in detail in section B4.7. With this perturbation the electric field at the end of the propagation through the material of length L becomes

$$\mathscr{E}(L,t) = \mathscr{E}_0 \cos\left(\omega t - \frac{2\pi}{\lambda_0} n_0 L - \frac{2\pi}{\lambda_0} KLv(t)\right) = \mathscr{E}_0 \cos(\omega t - mv(t)). \tag{B4.6}$$

Clearly, the output wave is phase modulated with a modulation index $m = (2\pi/\lambda_0)KL$. The term $(2\pi/\lambda_0)n_0L$ was dropped, because it is a fixed phase delay. In other words, by choosing the time origin appropriately, it can always be eliminated. The optical spectrum of such a phase-modulated wave can be quite complicated. For example, for a simple sinusoidal modulating voltage

$$\begin{aligned}\mathscr{E}(L,t) &= \mathscr{E}_0 \cos(\omega t + m \sin \omega_m t)\\ &= \mathscr{E}_0[J_0(m)\cos \omega_0 t + J_1(m)\cos[(\omega_0 + \omega_m)t] - J_1(m)\cos[(\omega_0 - \omega_m)t] + J_2(m)\cos[(\omega_0\\ &\quad + 2\omega_m)t] + J_2(m)\cos[(\omega_0 - 2\omega_m)t] + J_3(m)\cos[(\omega_0 + 3\omega_m)t] - J_3(m)\cos[(\omega_0 - 3\omega_m)t]\\ &\quad + J_4(m)\cos[(\omega_0 + 4\omega_m)t] - J_4(m)\cos[(\omega_0 - 4\omega_m)t] + \cdots]\end{aligned} \tag{B4.7}$$

where J_ms are the Bessel functions of the first kind of order m. As this expression shows, as soon as a single-frequency optical wave enters an index-modulated medium, its spectrum broadens and is no longer a single-frequency waveform due to phase modulation.

B4.4 Amplitude modulators

Amplitude modulation can be achieved based on absorption or index changes. Amplitude modulation based on index changes requires converting phase modulation into amplitude modulation. This is most commonly done using either Mach–Zehnder interferometers or directional couplers, which are described next.

B4.4.1 Directional coupler amplitude modulators

An optical directional coupler consists of two single-mode optical waveguides brought in close proximity over a length L as illustrated schematically in figure B4.2. If the separation between the waveguides is sufficiently small, the evanescent tail of the optical waveguide mode of one waveguide does not completely decay to zero before reaching the other waveguide. As a result, coupling between the waveguides occurs. Hence part of the wave in one waveguide can cross over or couple to the other waveguide. This situation is typically analysed using coupled mode theory [6, 12]. The field over the coupled section can be approximated as

$$\mathscr{E}(x,y,z) = a_1(z)e_1(x,y)e^{-j\beta_1 z} + a_2(z)e_2(x,y)e^{-j\beta_2 z}. \tag{B4.8}$$

This is the superposition of individual normalized waveguide modes, $e_1(x,y)$ and $e_2(x,y)$, with propagation constants β_1 and β_2. The amplitudes of the individual waveguide modes $a_1(z)$ and $a_2(z)$ are

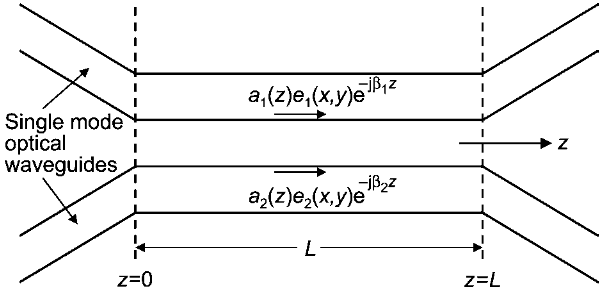

$a_1(z)e_1(x,y)e^{-j\beta_1 z}$
$a_2(z)e_2(x,y)e^{-j\beta_2 z}$
Single mode optical waveguides
z
L
z=0
z=L

Figure B4.2. Schematic of an optical directional coupler.

z dependent due to coupling between the waveguides. If only one of the waveguides is excited at the input, such that $a_1(0) = 1$ and $a_2(0) = 0$,

$$\mathscr{E}_1(x,y,L) = e_1(x,y)\left(\cos\left(\sqrt{\delta^2+\kappa^2}L\right) + \mathrm{j}\frac{\delta}{\sqrt{\delta^2+\kappa^2}}\sin\left(\sqrt{\delta^2+\kappa^2}L\right)\right)\mathrm{e}^{-\mathrm{j}\left(\frac{\beta_1+\beta_2}{2}\right)L} \tag{B4.9}$$

$$\mathscr{E}_2(x,y,L) = e_2(x,y)\left(\frac{\mathrm{j}\kappa}{\sqrt{\delta^2+\kappa^2}}\sin\left(\sqrt{\delta^2+\kappa^2}L\right)\right)\mathrm{e}^{-\mathrm{j}\left(\frac{\beta_1+\beta_2}{2}\right)L} \tag{B4.10}$$

where $\delta = (\beta_1 - \beta_2)/2$ is the detuning between the waveguides and κ is the coupling coefficient. κ depends on the degree of spatial overlap between the individual waveguide modes [6]. Hence over a given length a certain fraction of the input power will cross over to the other waveguide. In particular, if $\beta_1 = \beta_2 = \beta_0$

$$|a_1(z)|^2 = \cos^2(\kappa z) \text{ and } |a_2(z)| = \sin^2(\kappa z).$$

Hence for

$$\kappa L = \frac{\pi}{2} \text{ or } L = \frac{\pi}{2\kappa} \tag{B4.11}$$

complete cross over takes place. It is possible to eliminate complete cross over by modulating the index of refraction hence the propagation constant of one or both waveguides. For example if $\sqrt{\delta^2+\kappa^2}L = \pi$, all the optical power will switch back to the input waveguide. The required detuning is found after eliminating L using equation (B4.11) as

$$\delta = \frac{\beta_1 - \beta_2}{2} = \sqrt{3}\kappa. \tag{B4.12}$$

If we express β_1 and β_2 as

$$\beta_1 = \beta_0 + \Delta\beta_1 = \frac{2\pi}{\lambda_0}(n_0 + \Delta n_{1\mathrm{eff}}) = \frac{2\pi}{\lambda_0}(n_0 + K_1 v(t)) \tag{B4.13}$$

$$\beta_2 = \beta_0 + \Delta\beta_2 = \frac{2\pi}{\lambda_0}(n_0 + \Delta n_{2\mathrm{eff}}) = \frac{2\pi}{\lambda_0}(n_0 + K_2 v(t)) \tag{B4.14}$$

then

$$\delta = \frac{\pi}{\lambda_0}(K_1 - K_2)v(t) \tag{B4.15}$$

and the required external voltage is found as

$$V_\pi = \frac{\sqrt{3}\lambda_0 \kappa}{\pi(K_1 - K_2)} = \frac{\sqrt{3}\lambda_0}{2L(K_1 - K_2)}. \tag{B4.16}$$

The required voltage magnitude is minimized if $K_1 = -K_2 = K$. This requires index changes of equal magnitude and opposite sign in the coupled waveguides. This is known as the push–pull drive. Depending on the applied voltage, different degrees of coupling will occur. The variation of the coupled power as a function of the applied voltage is known as the transfer function of the modulator. Mathematically, this would be expressed in the cross over waveguide as

$$\frac{|a_2(L)|^2}{|a_1(0)|^2} = \frac{1}{1 + \left(\sqrt{3}\frac{v(t)}{V_\pi}\right)^2}\sin^2\left(\frac{\pi}{2}\sqrt{1 + \left(\sqrt{3}\frac{v(t)}{V_\pi}\right)^2}\right) \tag{B4.17}$$

where V_π is the required voltage swing to turn the modulator off and is given as

$$V_\pi = \frac{\sqrt{3}\lambda_0}{4LK}. \tag{B4.18}$$

The transfer function of a directional coupler modulator is shown in figure B4.3.

One can also use the result of this analysis to determine the chirp of a directional coupler modulator. Examining equation (B4.10) it is seen that the phase of $E_2(x, y, L)$ can be kept constant if the drive is push–pull, i.e. if β_1 increases a certain amount and β_2 decreases the same amount, which leaves $(\beta_1 + \beta_2)/2$ unchanged. This results in chirp free operation. This is not possible for $E_1(x, y, L)$ since its phase depends on the drive conditions. Therefore, if a chirp free operation is desired using a directional coupler modulator modulated power should be taken out of the cross over waveguide and push–pull drive should be used [7].

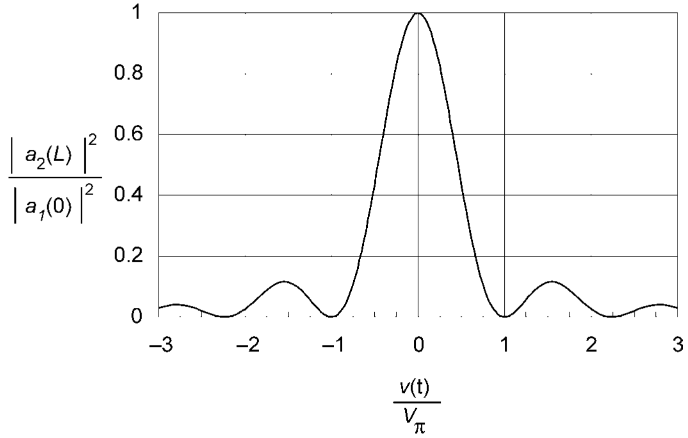


Figure B4.3. Transfer function of a directional coupler modulator.

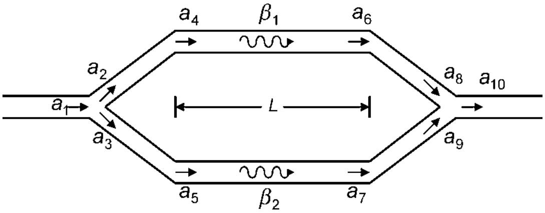


Figure B4.4. Schematic of a Mach–Zehnder interferometer.

B4.4.2 Mach–Zehnder amplitude modulators

Another commonly used amplitude modulator is a Mach–Zehnder interferometer shown in figure B4.4. This is the integrated optics version of the Michelson interferometer. The optical wave in the incoming waveguide is split into two equal parts using a Y-branch and propagates along the arms of the interferometer. At the other end waves in the two arms are combined into the output waveguide using another Y-branch. It is also possible to use 3 dB couplers in place of input and output Y-branches. For proper operation of the interferometer all the optical waveguides should be single mode. The mode amplitude in the output waveguide can be written as:

$$a_{10} = [s^2 e^{-j(\beta_1 L + 2\varphi_1)} + (1 - s^2) e^{-j(\beta_2 L + 2\varphi_2)}] a_1 \tag{B4.19}$$

where s^2 is the power splitting ratio in the Y-branch, φ_1 and φ_2 are the phase shifts along the arms of the Y-branch and β_1 and β_2 are the propagation constants of the arms of the interferometer. Any phase shift due to a length difference between the arms can also be included in φ_1 and φ_2 allowing us to assume that the arms have the same length L.

Ideally for equal power splitting $s^2 = 1/2$. However, due to imperfections during fabrication there could be slight deviations from the ideal value. In this case we can express s^2 as

$$s^2 = \frac{1}{2} + \xi.$$

Using this definition, equation (B4.19) can be expressed as

$$a_{10} = e^{-j\left(\left(\frac{\beta_1+\beta_2}{2}\right)L + (\varphi_1+\varphi_2)\right)} \left[\cos\left(\left(\frac{\beta_1 - \beta_2}{2}\right)L + (\varphi_1 - \varphi_2)\right) - 2j\xi \sin\left(\left(\frac{\beta_1 - \beta_2}{2}\right)L + (\varphi_1 - \varphi_2)\right)\right] a_1. \tag{B4.20}$$

The ratio of output power to input power is known as the transfer function of the Mach–Zehnder interferometer. Assuming that β_1 and β_2 can be perturbed with external voltages as shown in equations (B4.13) and (B4.14) and $\xi = 0$, and $\varphi_1 = \varphi_2$, the transfer function becomes

$$\frac{|a_{10}|^2}{|a_1|^2} = \cos^2\left(\frac{\pi}{\lambda_0}(K_1 - K_2) L v(t)\right). \tag{B4.21}$$

This transfer function is periodic with respect to the applied voltage. The modulator is turned off when

$$\frac{\pi}{\lambda_0}(K_1 - K_2) L v(t) = (\text{Odd integer}) \frac{\pi}{2}.$$

The lowest required voltage to turn the modulator off is known as V_π and is given as

$$V_\pi = \frac{\lambda_0}{2L(K_1 - K_2)}. \tag{B4.22}$$

Comparing this equation with equation (B4.16), we observe that the voltage swing required to turn a Mach–Zehnder modulator from on to off state is $\sqrt{3}$ times less than that required to turn a directional coupler modulator from on to off state for a given technology. Again, V_π is minimized if $K_1 = -K_2 = K$, i.e. for a push–pull drive. In terms of this definition of V_π, the transfer function becomes

$$\frac{|a_{10}|^2}{|a_1|^2} = \cos^2\left(\frac{\pi}{2}\frac{v(t)}{V_\pi}\right). \tag{B4.23}$$

The transfer function of a Mach–Zehnder modulator under push–pull drive is plotted in figure B4.5.

An important point in understanding the operation of a Mach–Zehnder modulator is physical explanation of what is happening in the off state. In the analysis, a lossless modulator is assumed. Equation (B4.21) predicts that there is no power output from the modulator in the off state even though there is input power. Since there are no reflections and optical absorption, one wonders what causes the power loss. This question can be answered with the help of figure B4.6, which describes the operation of a Mach–Zehnder modulator. In the on state, the incoming mode of the single-mode waveguide splits with equal amplitude to both arms of the interferometer in the input Y-branch. If $\beta_1 L = \beta_2 L$, the waves in both arms arrive at the output Y-branch with equal phase and combine to form the mode of the single-mode output waveguide. However in the off state $\beta_1 L - \beta_2 L = \pi$, and the waves in the both arms arrive at the output Y-branch with a π phase shift as shown in figure B4.6(*b*). As a result, when they are gradually combined by the Y-branch, in the output waveguide a mode with a null in the middle is excited. This would be the higher-order mode of the output waveguide. If the output waveguide is single mode, the higher-order mode would radiate out of the waveguide. As a result, there will not be any power left in the output waveguide after a certain length. This explanation shows that in the off state power is not lost. It simply radiates out of the output waveguide. This explanation also shows for proper operation the waveguides in the interferometer should be single mode, and the output waveguide should

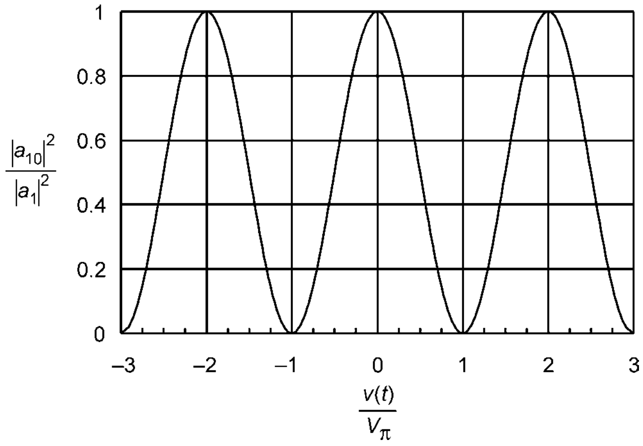


Figure B4.5. Transfer function of a Mach–Zehnder interferometer.

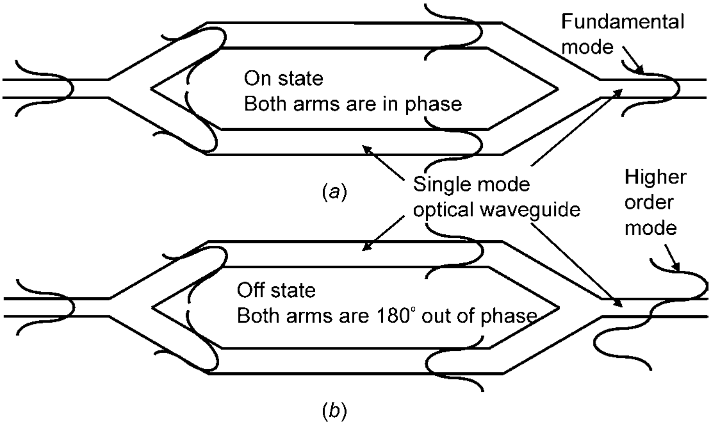


Figure B4.6. The physical description of a Mach–Zehnder modulator. (*a*) On state and (*b*) off state.

be sufficiently long for the radiation to take place in the off state. If the waveguides are not single mode in the off state only mode conversion between the fundamental and the higher-order mode takes place.

Examining equation (B4.20), it is observed that if $\xi = 0$ and $\varphi_1 = \varphi_2$

$$a_{10} = \mathrm{e}^{-\mathrm{j}\left(\frac{\beta_1+\beta_2}{2}\right)L}\left[\cos\left(\left(\frac{\beta_1-\beta_2}{2}\right)L\right)\right]a_1. \tag{B4.24}$$

If the drive is push–pull so that $(\beta_1 + \beta_2)/2$ is constant during modulation, the phase of the output signal does not change. Hence, chirp free operation results. Therefore, a push–pull driven Mach–Zehnder intensity modulator has no chirp.

B4.4.3 Electroabsorption amplitude modulators

In some materials the absorption at a given wavelength can be controlled using external voltages. Hence, it becomes possible to change the transmission through a waveguide with external voltages. This results in a very simple modulator, which is typically a short waveguide with an electroabsorption layer in it. However, critical control of the material composition and thickness are required. These devices are described in detail in section B4.7.1

B4.5 Travelling wave modulators

In optical modulators, the physical effects used to create index changes are very weak. As a result the K coefficients used in equations (B4.18) and (B4.22) are very small. This necessitates making the modulator electrode very long in order to have low drive voltages that can be supplied at high frequencies. Typical electrode lengths for electro-optic modulators are at the order of several centimetres. However, this increases the capacitance of the electrode drastically. The resulting RC product limits the bandwidth of modulation to less than a few gigahertz. These conflicting requirements on the electrode length can be eliminated using the travelling wave electrode concept. In this approach, the electrode is designed as a transmission line. Therefore, the electrode capacitance is distributed and does not create an RC limit on the modulator speed. A schematic of a travelling wave Mach–Zehnder modulator is shown in figure B4.7.

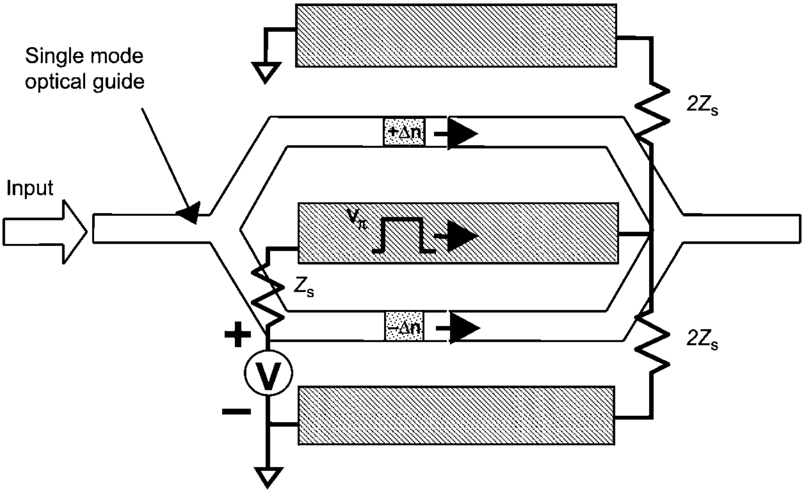


Figure B4.7. Schematic of a travelling wave Mach–Zehnder modulator.

In this figure, the electrode is designed as a coplanar waveguide. Electrical signal is applied from a voltage generator of internal impedance $Z_s\ \Omega$. In this case, signal is an electrical pulse. Both the electrical pulse and the CW optical signal travel in the same direction. On the part of the electrode where the voltage is present the index of the material is modified. In this case a push–pull drive is considered, hence index is increased a certain amount in one arm of the interferometer and decreased the same amount on the other arm. Clearly if the optical signal travels at the same velocity as the electrical signal it will experience the same index change all along the electrode. Hence, the phase shift it experiences will be integrated over the electrode length. Therefore, a very small index change may create a large phase shift. Any velocity mismatch between the electrical and the optical signals will reduce the net phase shift. The small signal modulation response of a travelling wave modulator whose electrode is terminated by its characteristic impedance is given as [8]

$$M(f) = \mathrm{e}^{-\left(\frac{\alpha_m L}{2}\right)} \left[\frac{\sinh^2\left(\frac{\alpha_m L}{2}\right) + \sin^2\left(\frac{\pi f (n_\mu - n_o) L}{c}\right)}{\left(\frac{\alpha_m L}{2}\right)^2 + \left(\frac{\pi f (n_\mu - n_o) L}{c}\right)^2} \right]^{\frac{1}{2}} \tag{B4.25}$$

where α_m and L are the loss coefficient and the length of the electrode, respectively. f is the electrical frequency and c is the speed of light in vacuum. n_μ and n_o are the microwave and optical indices and are related to the microwave and optical velocities through well-known expressions. The electrical waveform applied to the electrode of the modulator has a certain shape and contains many frequency components. The CW light that enters the modulator also becomes phase modulated as soon as it starts to interact with the electrical signal and exhibits a spectral range. As a result, the centre of gravity of both the electrical and optical signals should match [9]. That requires group velocity matching, hence the microwave and optical indices defined above should be the group indices. Usually, the modulator electrode is a quasi-TEM transmission line and has no or very little dispersion. That makes the group and the phase velocities on the electrode the same. If the electrode exhibits significant dispersion complications may arise. For example if the electrode is driven by a CW microwave signal, as in a pulse generation application, microwave phase and optical group velocities should be matched. However, if a digital electrical signal is applied to the electrode, optical and microwave group velocities should be matched. This can create velocity mismatch depending on the application. Furthermore, optical group velocity may change due to material dispersion. For example, a modulator velocity matched at 1.5 μm can be mismatched at 1.3 μm [10, 11]. This is especially a problem for semiconductor modulators, where material dispersion is much more significant compared to $LiNbO_3$.

Even in the case of perfect velocity matching, bandwidth is limited by the microwave loss of the electrode. At high microwave frequencies the microwave loss increases, which reduces the voltage on the line. Hence modulation is no longer as effective as it was at lower frequencies where microwave loss was lower. Based on equation (B4.25) if there is no velocity mismatch 3 dB bandwidth will be at the frequency where the total electrode loss becomes 6.34 dB. Therefore, a low loss and velocity matched electrode is essential for the realization of a very wide-bandwidth travelling wave modulator. Another very important consideration is impedance matching the modulator electrode. If the electrode is terminated with impedance having a value different than the electrode impedance, there is an impedance mismatch at the end of the modulator. This mismatch creates a reflected voltage wave travelling in the opposite direction. The reflected wave interferes with the wave travelling in the same direction as the optical wave and a standing wave is formed on the electrode. As a result, voltage on some parts of the electrode is higher than the rest. These parts modulate more effectively, whereas modulation is less effective on the parts having less voltage amplitude. Variation of the standing wave voltage is frequency dependent and the modulation efficiency could significantly vary as a function of frequency. In modulators with long electrodes the counter-propagating part of the standing wave is badly mismatched and contributes to modulation only at low frequencies. In these cases, modulation efficiency drops rapidly over a few gigahertz.

Based on this discussion the following requirements should be satisfied to take full advantage of the travelling wave idea:

(a) Propagation loss of the optical guide should be low so that a long modulator can be realized. This helps to significantly reduce the drive voltage.

(b) The microwave electrode mode should be quasi-TEM hence the phase and the group velocities are the same for the electrode.

(c) Microwave and optical group velocities should be matched.

(d) The electrode microwave loss coefficient should be low, so that a long modulator can be realized. As described before total electrode loss should be lower than 6.34 dB at the desired 3 dB bandwidth point.

(e) The electrode should be terminated by its characteristic impedance so that there is no standing wave on the electrode. It is desirable to have an electrode characteristic impedance of 50 Ω, but this may not always be possible.

B4.6 Physical effects used in optical modulators

The physical mechanisms used in optical modulators are change in absorption and the index of refraction under external perturbations. Change in absorption and the index of refraction are fundamentally related. The most commonly used external perturbation is an externally applied electric field, which can be generated electronically by applying voltages to the material. Modulators utilizing index changes and absorption changes under an external electric field are known as electro-optic and electroabsorption modulators, respectively. Index and absorption changes can also be generated using pressure or strain and magnetic fields. The strain is typically generated using acoustic waves.

Modulators using acoustic waves to induce index changes due to strain are known as acousto-optic modulators. Similarly, modulators using magnetic fields to induce index changes due to magneto-optic effect are known as magneto-optic modulators. Absorption is most important in III–V compound semiconductors, such as GaAs, AlAs, InP, and their alloys AlGaAs and InGaAsP.

B4.6.1 Change of absorption in semiconductors under external electric fields

In semiconductors, the sources of optical absorption are band-to-band transitions, free carriers, intraband transitions, excitons, transitions between band tails, transitions between bands and impurities, and acceptor to donor transitions. Of these band to band absorption, free carrier absorption, intraband absorption and excitonic absorption are the most dominant processes in good quality material. These processes can be perturbed by external electric fields. They are briefly described below.

Band to band absorption and Franz–Keldysh effect

In semiconductors, an incoming photon can excite an electron from the valence band to the conduction band under certain conditions as shown in figure B4.8. Since the photon momentum h/λ, where h is Planck's constant and λ is the wavelength of the light, is much smaller than the crystal momentum h/b, where b is the lattice constant, momentum conservation requires that the momentum of the initial and final states remain unchanged. Hence, only vertical transitions are allowed. Furthermore, in order to absorb a photon the initial state in the valence band should be filled and the final state in the conduction band should be empty. The energy of the photon should be larger than the energy difference between the filled initial state and the empty final state of the semiconductor. This energy difference is typically the bandgap of the material, but depending on the doping type and level, the required photon energy could be larger than the bandgap. The absorption coefficient α due to band-to-band absorption in a direct bandgap semiconductor can be expressed as

$$\alpha = A|R(h\nu)|^2 \rho_r(h\nu)(F_1 - F_2)$$

where R is the matrix element, $h\nu$ is the photon energy, F_1 and F_2 are the Fermi factors for the valence and conduction bands, respectively, ρ_r is the reduced density of states and A is a coefficient that depends

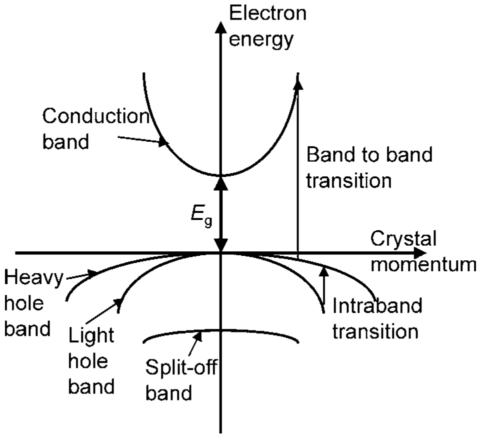


Figure B4.8. Band structure of a typical III–V compound semiconductor in the vicinity of the zone centre. Various possible transitions responsible for dominant optical absorption are also illustrated.

on fundamental constants [12]. The matrix element R is proportional to the overlap of the electronic wave functions in the conduction and valence bands. The value of R depends on the material type and its dimensionality and the polarization of the optical wave. Polarization dependence typically appears in lower-dimensional systems such as quantum wells (QWs). ρ_r also depends on the material dimensionality. For bulk materials absorption can be expressed as [12]

$$\alpha = B(h\nu - E_g)^{\frac{1}{2}}$$

assuming $F_1 = 1$ and $F_2 = 0$ as in undoped bulk material. In this expression, E_g is the bandgap of the material.

In the presence of an electric field energy bands are tilted. Figure B4.9 shows the variation of the conduction band minimum and valence band maximum in space for different values of applied electric field. In the absence of an external electric field and if the material properties such as doping and composition are uniform band extremes are flat. Increasing external electric field values tilt the energy bands, which increases the penetration of the electron and hole wave functions into the bandgap as shown in figure B4.9. This in turn increases the overlap of the electron and holes in the bandgap. In figure B4.9, points A and B are the classical turning points for the electron and hole wave functions. In other words at these points the wave functions change from oscillatory to exponentially decaying behaviour. Due to penetration of the wave functions into the bandgap it is possible to excite an electron from the valence band to conduction band with a photon of energy $h\nu$ as seen in figure B4.9. This energy is clearly less than the bandgap of the material. Hence, it becomes possible to absorb at photon energies less than the bandgap energy and this absorption increases with increasing external field. As a result, the absorption tail extends to shorter energies or longer wavelengths. This phenomenon is known as the Franz–Keldysh effect [13, 14]. It is possible to modulate the absorption from low to high absorption in

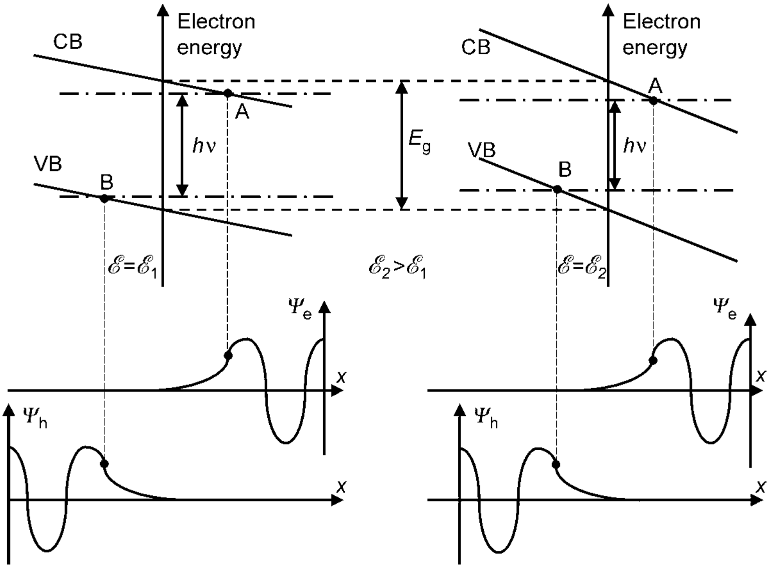


Figure B4.9. Variation of the conduction band minimum and valence band maximum under different applied external electric fields for a material with uniform doping and composition. Penetration of the electron and hole wavefunctions (Ψ_e and Ψ_h) into the bandgap is also illustrated.

the long-wavelength side with external fields. This situation is similar to tunnelling through a potential barrier under an applied electric field. In this case increasing external electric field effectively reduces the width of the barrier due to more and more penetration of the wave functions into the bandgap. The absorption coefficient under an external applied field, α_{E}, can be expressed as [15, 16]

$$\alpha_{\mathrm{E}} = B\sqrt{h\nu_{\mathrm{E}}}\pi\int_{\frac{E_{\mathrm{g}}-h\nu}{h\nu_{\mathrm{E}}}}^{\infty} \mathrm{A}i(x)^2 \mathrm{d}x$$

where

$$\nu_{\mathrm{E}} = \left(\frac{e^2\mathscr{E}^2\hbar^2}{2m_{\mathrm{r}}}\right)^{\frac{1}{3}}.$$

In this equation Ai(x) is the Airy function and m_{r} is the reduced effective mass. The absorption is also modified for photon energies larger than the bandgap energy. In this case, absorption becomes oscillatory since above the bandgap electron and hole wave functions are oscillatory. Therefore, bringing them closer changes their overlap in an oscillatory manner. Below the bandgap the wave functions were exponentially decaying, hence bringing them closer increased their overlap monotonically. The absorption changes above the bandgap are not used for modulation due to large background absorption.

Excitonic absorption in bulk and quantum well material and the quantum confined Stark effect

In a material free electrons and holes attract one another and form a bond similar to a hydrogen atom. This is called an exciton. An exciton has a strong absorption somewhat similar to an atomic absorption. In an exciton electron–hole interaction is treated as a Coulomb interaction between two point charges. Using the hydrogen atom model the binding energy for an exciton can be expressed in electron volts as $E_{\mathrm{x}n} = (13.6/n_{\mathrm{x}}^2)\ (m_{\mathrm{r}}/\varepsilon_{\mathrm{r}}^2)$ where n_{x} is an integer representing different energy levels, i.e. the quantum number, m_{r} is the reduced effective mass, and ε_{r} is the relative dielectric constant. Another important parameter is the exciton dimension. Again using the hydrogen atom model exciton radius can be expressed in Å as $a_{\mathrm{x}n} = 0.53(\varepsilon_{\mathrm{r}}/m_{\mathrm{r}})n_{\mathrm{x}}^2$. In a bulk compound semiconductor, $E_{\mathrm{x}1}$ is about 5 meV and $a_{\mathrm{x}1}$ is about 150 Å. As a result a bulk exciton covers many lattice sites and is weakly bound. Therefore, it is observed only at low temperatures. However, in a QW, electrons and holes are confined in the same physical space. As a result, they overlap and interact strongly and form an exciton confined in the QW. This strong confinement increases the binding energy or the energy required to ionize an exciton into an electron–hole pair. For a purely two-dimensional exciton the increase in the binding energy would be a factor of four [17]. However, in a QW the exciton wave function penetrates into the barriers and the dimensionality of the exciton is somewhere in between two and three. For a very thin QW this penetration could be excessive and the exciton behaves like a three-dimensional exciton. The increase in the binding energy due to quantum confinement makes the binding energy of the exciton larger than the broadening due to phonon scattering at room temperature. As a result, excitonic absorption is observed at room temperature. Spectra of such strong excitonic absorption are very sharp, and are localized in the vicinity of wavelengths corresponding to the bandgap of the QW. This energy is larger than the bandgap of the QW material due to quantization of electron and hole energy levels as shown in figure B4.10. When an external electric field is applied the electron and hole are forced to opposite ends of the QW and are physically separated as seen in figure B4.10. The applied field shifts the energy levels in the QW. This situation is similar to the shift in the energy levels of an atom under an applied electric field, which is known as the Stark effect. Hence, the corresponding transition energy shift in a QW is known as the quantum confined Stark effect (QCSE). In the QCSE, the spatial overlap or interaction of the electron

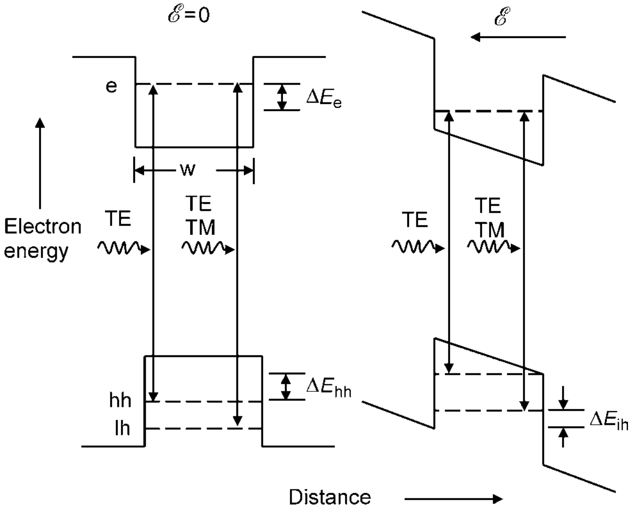


Figure B4.10. Schematic of an unstrained QW energy diagram with and without an applied electric field ℰ. e, hh, and lh designate energy levels for electrons, heavy holes and light holes, respectively. After [1] figure 3. Reproduced by permission of IEEE.

and the hole is reduced and excitonic absorption is decreased and broadened. This makes it possible to modulate the absorption very strongly with external fields around a narrow wavelength range.

Figure B4.11 shows the photocurrent spectra of an unstrained MQW material as a function of wavelength at different applied voltages [18], where device photocurrent is proportional to optical absorption. Two peaks are resolved in the absorption spectra. These are due to excitons formed between electrons and heavy holes (hh) and electrons and light holes (lh). Transition energies of hh and lh excitons are different due to different effective mass of lh and hh as seen in figure B4.11. Furthermore, the excitons interact with different optical polarizations. The hh excitons interact with TE polarized light and lh excitons interact with both TE and TM polarized light. In figure B4.11, the first absorption peak has a lower transition energy or higher optical wavelength and corresponds to the lowest hh exciton indicated in figure B4.10. The second absorption peak corresponds to the lh exciton. As bias voltage increases absorption characteristics broaden and peak absorption decreases and moves towards longer

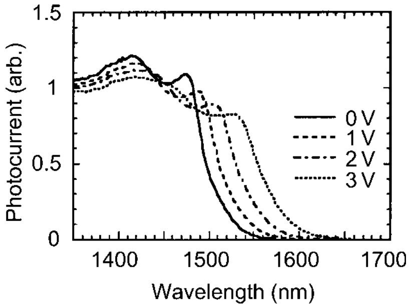


Figure B4.11. Photocurrent spectra of an unstrained MQW modulator as a function of bias voltage. The incident light is TE polarized. After [30] figure 2. Reproduced by permission of IEEE.

wavelengths. For example, at 1.55 μm absorption is modulated strongly when bias changes between 0 and 3 V. For low bias voltages, the shift in the ground state energy ΔE_{i} of a particle in a QW can be approximated as [19]

$$\Delta E_{\mathrm{i}} = C m_{\mathrm{i}} w^4 \mathscr{E}^2 \quad \mathrm{i} = \mathrm{e}, \mathrm{hh}, \mathrm{lh} \tag{B4.26}$$

where m_{i} is the effective mass of the particle, w is the width of the QW and C is a constant. Hence, wide QWs are desired for efficient operation.

Free carrier absorption

Free carriers are electrons and holes free to move in the conduction and valence bands. They can interact with photons and make transitions to a higher energy in the same band by absorbing a photon. This is known as an intra-band absorption. Since such transitions are not vertical, they require additional interactions to conserve momentum. The required momentum change can be provided by phonons or scattering from ionized impurities. There is no energy threshold for this transition; its spectrum is typically monotonic and covers a very wide range. Using a simple Drude model, which models an electron in an oscillatory field as a damped oscillator, absorption coefficient due to free carriers, α_{f}, can be expressed as [20]

$$\alpha_{\mathrm{f}} = \frac{N q^2 \lambda^2}{8 \pi^2 m^* n c^3 \tau} \tag{B4.27}$$

where N is the electron concentration, τ is the carrier relaxation time, m^* is the effective mass, n is the index of refraction, c is the speed of light in vacuum and λ is the free space wavelength. Other formulations in which the momentum is provided by acoustic phonons, optical phonons or ionized impurities yield a wavelength dependence of λ^p, where p can range from 1.5 to 3.5 [20]. Since one of the elements that affect τ is ionized impurity concentration, its value depends on the doping level and type. At high doping concentrations the carrier concentration dependence of α_{f} could be more like $N^{3/2}$. Around 1.55 μm, $\alpha_{\mathrm{f}} = 1 \times 10^{-18} N\,(\mathrm{cm}^{-3})$ in GaAs [21]. The situation is similar for intra-band absorption involving holes. However, there is another source of absorption due to possibility of transition between light and heavy hole bands in p type material. The most likely source for this absorption is vertical or near vertical transitions at longer wavelengths. Such absorption is known as intra-band absorption and is indicated schematically in figure B4.8.

Electro-optic effects

The real and imaginary parts of a complex function are related to one another if it has no poles in the lower or upper complex plane. If one describes the index of refraction of a material as a complex function, the real and imaginary parts will be related. In this representation, the real and imaginary parts of the complex index of refraction are the index of refraction and the absorption of the material. As a result, a change in absorption will generate a change in the refractive index and vice versa. This relationship is known as the Kramers–Kronig relation and can be expressed as

$$\Delta n(\mathscr{E}', \mathscr{E}) = \frac{hc}{\pi} \int_0^\infty \frac{\Delta\alpha(\mathscr{E}', \mathscr{E})}{\mathscr{E}'^2 - \mathscr{E}^2} \,\mathrm{d}\mathscr{E}'. \tag{B4.28}$$

Hence any effect that creates an absorption change $\Delta\alpha$ under an applied electric field $\mathscr{E}$ at photon energy E will also create an index change Δn. These index changes are known as electro-optic effects and are described below.

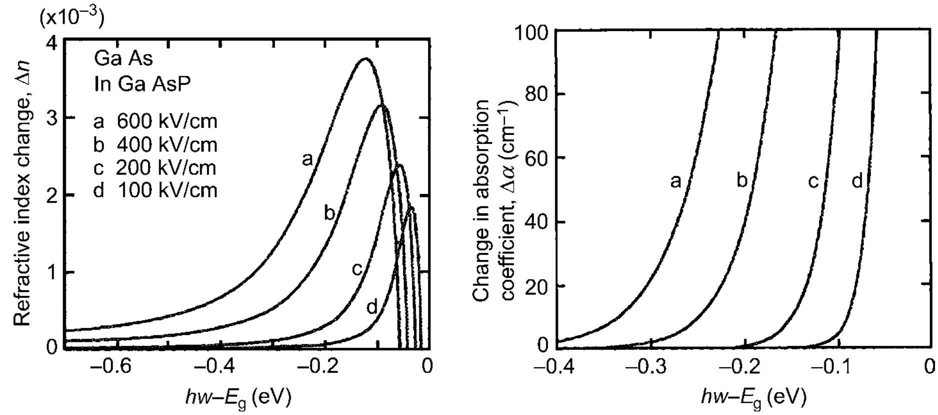


Figure B4.12. The calculated change in the index of refraction and absorption coefficient as a function of photon energy for GaAs and InGaAsP at different electric field strengths. After [22] figures 1 and 5. Reproduced by permission of AIP.

Electrorefractive effect

The electrorefractive effect is the index change accompanied with Franz–Keldysh and QCSE. Electrorefraction in bulk GaAs and InGaAsP was calculated in [22]. The calculated change in the index of refraction and absorption coefficient as a function of photon energy for GaAs and InGaAsP at different electric field strengths are shown in figure B4.12.

Index change increases rapidly towards the bandgap. The peak of the increase shifts away from the band edge with increasing electric field. For wavelengths very close to the bandgap index change decreases and it should eventually be negative. Figure B4.13 shows the change in the refractive index in

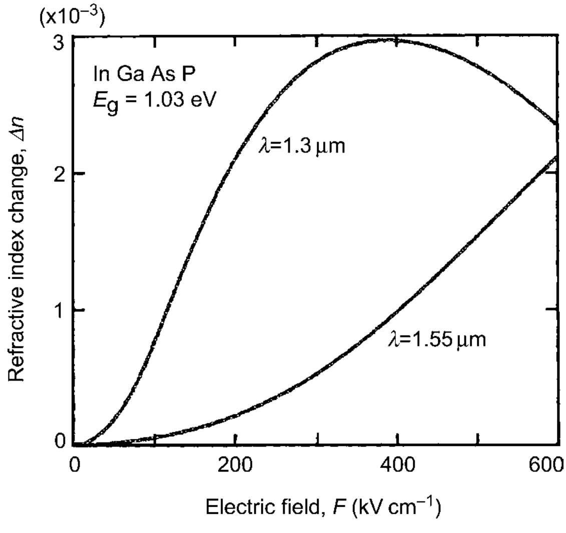


Figure B4.13. The change in the refractive index Δn in $In_{0.76}Ga_{0.24}As_{0.52}P_{0.48}$ as a function of the applied electric field at two different wavelengths. After [22] figure 3. Reproduced by permission of AIP.

$In_{0.76}Ga_{0.24}As_{0.52}P_{0.48}$ as a function of the applied electric field at two different wavelengths. For photon energies far below the bandgap the refractive index for a given electric field can be expressed as

$$\Delta n = G(\lambda)\mathscr{E}^2. \tag{B4.29}$$

There is a quadratic dependence on the electric field and the wavelength dependent coefficient $G(\lambda)$ can be expressed for GaAs as

$$G(\lambda) = 3.45 \times 10^{-16} \exp\left(\frac{3}{\lambda^3}\right)\ \mathrm{cm^2\,V^{-2}} \tag{B4.30}$$

where wavelength is expressed in micrometres [23]. For InGaAsP, $G(1.55\,\mu\mathrm{m}) = 5.8 \times 10^{-15}\,\mathrm{cm^2\,V^{-2}}$.

At high electric fields it is possible to raise the refractive index by about 3×10^{-3} at photon energies near the bandgap. However, this index change is accompanied by a large absorption change due to the Franz–Keldysh effect. In order to get predominantly an electro-optic effect absorption change should be kept to a minimum. The relative variation of the real and imaginary parts of the index of refraction with applied field is quantified using the figure of merit $\Delta n/\Delta k$, where Δk is the change in the imaginary part of the refractive index. Δk and $\Delta\alpha$ are related as $\Delta\alpha = 4\pi\Delta k/\lambda$. As described earlier, it is possible to modify the absorption in a QW close to the band edge strongly due to QCSE. This also creates an accompanying index change. Δn and $\Delta n/\Delta k$ as functions of the electric field for two MQW samples at two different wavelengths are shown in figure B4.14 [24]. In both cases, Δn shows a quadratic dependence on the applied field. $G(1.537\,\mu\mathrm{m}) = 6.73 \times 10^{-13}\,\mathrm{cm^2\,V^{-2}}$ for a 85 Å InGaAsP QW of bandgap energy 1.57 μm within 85 Å InP barriers. $G(1.306\,\mu\mathrm{m}) = 7.32 \times 10^{-13}\,\mathrm{cm^2\,V^{-2}}$ for a 70 Å InGaAsP QW of bandgap energy 1.33 μm within 250 Å InP barriers [24]. These values are more than two orders of magnitude larger than the corresponding bulk material values. But this does not necessarily make a much better modulator since the overlap of the optical mode with the QW material in a typical modulator is also a few per cent. However, the spectral width of the absorption in a QW material is much narrower compared to bulk material. Furthermore QCSE red shifts this relatively narrow resonance, hence for a given wavelength detuning $\Delta n/\Delta k$ value is larger in the QW material compared to bulk material.

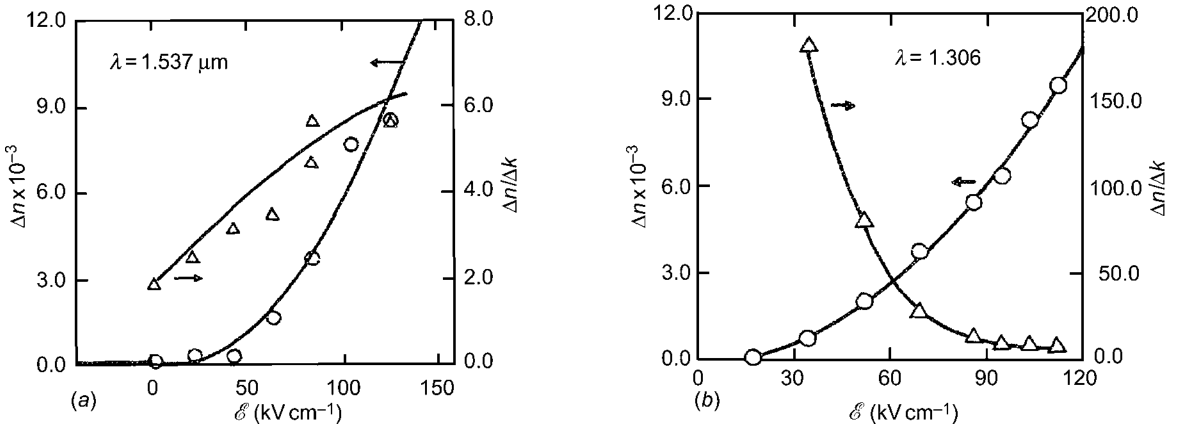


Figure B4.14. Change in the refractive index, Δn, and the ratio of the changes in the real and imaginary parts of the refractive index, $\Delta n/\Delta k$, as functions of the electric field for two MQW samples: (*a*) ten periods of 85 Å InP barriers and 85 Å InGaAsP wells of bandgap energy 1.57 μm at 1.537 μm and (*b*) five periods of 250 Å InP barriers and 70 Å InGaAsP wells of bandgap energy 1.33 μm at 1.16 μm. After [24] figure 3. Reproduced by permission of AIP.

Plasma effect

The plasma effect is the accompanying index change associated with the free carrier absorption. For n-type GaAs of carrier concentration N the index change Δn_{N} with respect to undoped material is

$$\Delta n_{\mathrm{N}} = -9.6 \times 10^{-21} \frac{N}{nE^2}$$

where n is the index of refraction and E is the photon energy. For p-type GaAs of carrier concentration P the corresponding index changes for intra- and inter-band transitions are

$$\Delta n_{\mathrm{PInter}} = -6.3 \times 10^{-22} \frac{P}{E^2} \quad \text{and} \quad \Delta n_{\mathrm{PIntra}} = -1.8 \times 10^{-21} \frac{P}{nE^2}.$$

As these expressions indicate removal of electrons or holes in a doped semiconductor reduces free carrier absorption, which in turn increases the index of refraction.

Band filling effect

In doped material, the position of the Fermi level moves depending on the doping level. In very heavily doped material, the Fermi level can move close to or even into the conduction or valence bands. Therefore, the absorption threshold could be different between the doped and undoped material. Therefore, at photon energies near the bandgap absorption could be stronger for intrinsic material compared to heavily doped material. This creates an index change due to Kramers–Kronig relationship between doped and undoped materials. This index change can be expressed as [23, 25]

$$\Delta n(E) = H(E)N.$$

The coefficient H strongly peaks around photon energy E corresponding to the bandgap of the material [23]. Below the bandgap its value is about $5 \times 10^{-21}\,\mathrm{cm}^{-3}$ for GaAs. This relationship holds as long as the semiconductor is not very heavily doped and the doping level is less than about $5 \times 10^{17}\,\mathrm{cm}^{-3}$. For heavier doping band tails that develop at the conduction band edge and bandgap shrinkage start to dominate the absorption. This increase in the absorption starts to give an index change that opposes the band filling effect [26]. For p-type material the variation of the Fermi level with doping concentration is much less due to heavier hole effective mass. Hence the band-filling effect is much weaker in p-type material.

Linear electro-optic effect

The linear electro-optic effect can be considered as the accompanying index change due to absorption at ultraviolet wavelenghts. However, the tail of absorption at such high photon energies is very weak around 1.55 μm. Hence this effect is considered as a pure index change and modelled in a different way. In a dielectric material, electric field $\vec{\mathrm{E}}$ and electric flux density $\vec{\mathrm{D}}$ are related through the dielectric tensor, which can be expressed as:

$$\begin{bmatrix} \mathcal{E}_x \\ \mathcal{E}_y \\ \mathcal{E}_z \end{bmatrix} = \frac{1}{\varepsilon_0} \begin{bmatrix} \left(\frac{1}{n^2}\right)_{xx} & \left(\frac{1}{n^2}\right)_{xy} & \left(\frac{1}{n^2}\right)_{xz} \\ \left(\frac{1}{n^2}\right)_{xy} & \left(\frac{1}{n^2}\right)_{yy} & \left(\frac{1}{n^2}\right)_{yz} \\ \left(\frac{1}{n^2}\right)_{xz} & \left(\frac{1}{n^2}\right)_{yz} & \left(\frac{1}{n^2}\right)_{zz} \end{bmatrix} \begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix}. \tag{B4.31}$$

The index of refraction of the crystal n is defined as $n = \sqrt{\varepsilon_{\mathrm{r}}}$. This tensor has only six independent components [27]. The electric energy density W_e in a medium is given as $W_e = (1/2)\vec{E}\cdot\vec{D}$. Using the relationship shown in equation (B4.31) this can be written as

$$2\varepsilon_0 W_e = \left(\frac{1}{n^2}\right)_{xx} D_x^2 + \left(\frac{1}{n^2}\right)_{yy} D_y^2 + \left(\frac{1}{n^2}\right)_{zz} D_z^2 + 2\left(\frac{1}{n^2}\right)_{yz} D_y D_z + 2\left(\frac{1}{n^2}\right)_{xz} D_x D_z + 2\left(\frac{1}{n^2}\right)_{xy} D_x D_y. \tag{B4.32}$$

It should be noted that in equation (B4.32) $\left(1/n^2\right)_{xx} \neq \left(1/n_{xx}^2\right)$ unless the principal axes are used as the coordinate system. Since the left hand side of equation (B4.32) is a constant, the following association can be made:

$$x = \frac{D_x}{\sqrt{C}}, \quad y = \frac{D_y}{\sqrt{C}}, \quad \text{and} \quad z = \frac{D_z}{\sqrt{C}},$$

where

$$C = \sqrt{2\varepsilon_0 W_e}.$$

With this association and using the abbreviated notation in which $xx \equiv 1$, $yy \equiv 2$, $zz \equiv 3$, $yz \equiv 4$, $xz \equiv 5$, and $xy \equiv 6$, equation (B4.32) can be rewritten as

$$\left(\frac{1}{n^2}\right)_1 x^2 + \left(\frac{1}{n^2}\right)_2 y^2 + \left(\frac{1}{n^2}\right)_3 z^2 + 2\left(\frac{1}{n^2}\right)_4 yz + 2\left(\frac{1}{n^2}\right)_5 xz + 2\left(\frac{1}{n^2}\right)_6 xy = 1. \tag{B4.33}$$

In this equation x, y and z relate to the polarization of the optical field. For example, $y = z = 0$ means an x polarized optical field. Equation (B4.33) represents an ellipse and is known as the index ellipsoid or optical indicatrix. Using this equation one can find the phase velocity of an optical wave propagating in the crystal in any arbitrary direction [27].

If an external electric field is applied to the material, the coefficients of the index ellipsoid change. This external electric field is the modulating field and typically extends in frequency from DC to the millimetre wave range. This change in the coefficients is equivalent to modulating the velocity of the optical wave and is used to create the modulation of the optical wave. This change is linearly proportional to the applied field; hence the relationship between the six coefficients in equation (B4.33) and the three electric field components can be expressed as

$$\begin{pmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} \mathscr{E}_x \\ \mathscr{E}_y \\ \mathscr{E}_z \end{pmatrix}. \tag{B4.34}$$

The 6×3 matrix multiplying the electric field vector is known as the electro-optic tensor and its elements r_{ij} are known as the electro-optic coefficients. The typical magnitude of these coefficients is of the order of $10^{-12}\,\mathrm{m\,V^{-1}}$. With these changes the index ellipsoid is written as

$$\underset{(a)}{\begin{bmatrix} 0 & -r_{22} & r_{13} \\ 0 & r_{22} & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{51} & 0 \\ r_{51} & 0 & 0 \\ -r_{22} & 0 & 0 \end{bmatrix}} \qquad \underset{(b)}{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & r_{41} & 0 \\ 0 & 0 & r_{41} \end{bmatrix}}$$

Figure B4.15. Electro-optic tensor (*a*) for 3*m* crystal symmetry, and (*b*) for cubic crystal symmetry.

$$(x \quad y \quad z)\begin{bmatrix} \left(\left(\frac{1}{n^2}\right)_1+\Delta\left(\frac{1}{n^2}\right)_1\right) & \left(\left(\frac{1}{n^2}\right)_6+\Delta\left(\frac{1}{n^2}\right)_6\right) & \left(\left(\frac{1}{n^2}\right)_5+\Delta\left(\frac{1}{n^2}\right)_5\right) \\ \left(\left(\frac{1}{n^2}\right)_6+\Delta\left(\frac{1}{n^2}\right)_6\right) & \left(\left(\frac{1}{n^2}\right)_2+\Delta\left(\frac{1}{n^2}\right)_2\right) & \left(\left(\frac{1}{n^2}\right)_4+\Delta\left(\frac{1}{n^2}\right)_4\right) \\ \left(\left(\frac{1}{n^2}\right)_5+\Delta\left(\frac{1}{n^2}\right)_5\right) & \left(\left(\frac{1}{n^2}\right)_4+\Delta\left(\frac{1}{n^2}\right)_4\right) & \left(\left(\frac{1}{n^2}\right)_3+\Delta\left(\frac{1}{n^2}\right)_3\right) \end{bmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 1. \tag{B4.35}$$

Crystal symmetry imposes certain restrictions on the form of the index ellipsoid. Figure B4.15 shows the form of the electro-optic tensor for 3*m* and cubic crystal symmetries. More complete representation of the electro-optic tensor for other crystal symmetries can be found in [28].

B4.7 Specific examples of different modulator technologies

B4.7.1 Electroabsorption modulators

Lumped electroabsorption modulators

Earlier it was shown that in a QW it was possible to modulate the absorption very strongly with external fields around a narrow wavelength range due to QCSE. If such QWs are embedded in an optical waveguide, the insertion loss of this waveguide can be modulated by applying an electric field and changing the absorption of the QWs through the QCSE [29]. Figure B4.16 shows such a modulator [1, 30]. Typically, MQWs are used to increase absorption and are embedded in the i region of a reverse

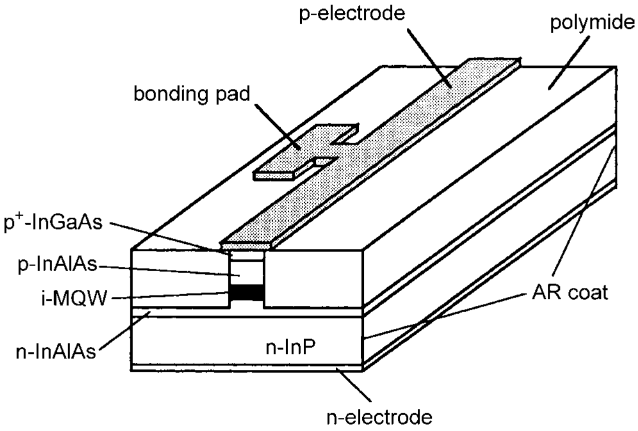


Figure B4.16. Schematic of an MQW electroabsoption modulator. After [30] figure 1. Reproduced by permission of IEEE.

biased p-i-n diode. The typical thickness of the i region, d_i, is in the 0.1–0.5 μm range. Therefore, it is possible to apply very strong electric fields with a few volts of reverse bias. In this case, well and barrier material are InGaAs and InAlAs, respectively. The index of refraction of the top p and bottom n-InAlAs is lower than that of the MQW. This forms a slab waveguide. By deeply etching this slab waveguide a channel optical guide is formed. Waveguide widths are of the order of 1–3 μm and the etch depths are about 1–2 μm. Typical device lengths are in the 50–300 μm range. Ohmic contacts are formed at the top and bottom. Etched areas underneath the bonding pads are filled with a low-dielectric-constant polyimide to reduce device capacitance. Since only the absorption modulation is utilized, the optical waveguide need not be single mode. Therefore, the profile of the optical waveguide can be optimized for improved coupling to the incoming fibre mode. However, critical control of the composition and thickness of the epitaxial layers are required.

The transmission through such a modulator as a function of applied voltage can be expressed as

$$T(V) = T_0 \exp[-\Gamma \alpha(V) L] \tag{B4.36}$$

where Γ is the overlap between the optical mode and the MQW, $\alpha(V)$ is the absorption coefficient as a function of applied voltage V, L is the length and T_0 is the coupling coefficient between the fibre and the optical waveguide. The on/off ratio of an EA modulator in decibels can be expressed as

$$10\log\left[\frac{T(V)}{T(0)}\right] = \frac{\alpha(V)-\alpha(0)}{\alpha(0)} 10\log\left[\mathrm{e}^{-\alpha(0)L}\right] = \frac{\Delta\alpha}{\alpha}[\text{Propagation loss (dB)}]. \tag{B4.37}$$

Optical propagation loss of EA modulators is large, typically in the 15–20 dB mm^{-1} range. Main components of this loss are the free carrier absorption, especially in the p layers, and band-to-band absorption, both described in earlier sections. The second loss component can be made smaller by increasing the separation between the wavelength of operation and the absorption peak, which is called detuning. Typical detuning values are about 20–50 nm. Typical $\Delta\alpha/\alpha$ values are in the 3–10 range. Large-on/off-ratio devices can be obtained using long devices, but that also increases the insertion loss. For the typical EA modulator lengths in the 50–300 μm range propagation loss is 1–3 dB. Therefore, to get large extinction ratios with low device insertion loss, $\Delta\alpha/\alpha$ should be maximized. Furthermore, efficient modulation requires a large $\Delta\alpha/\Delta v$ or $(1/d_i)(\Delta\alpha/\Delta\mathscr{E})$, where $\mathscr{E}$ is the applied electric field. In other words, large Stark shifts are required. For low bias voltages, the shift in the ground state energy ΔE_i of a particle in a QW is proportional to the fourth power of the well width as shown in equation (B4.26) [19]. Hence wide QWs are desired for efficient operation.

For fibre optic communication applications the desired wavelength is around 1.55 μm. This requires alloys of InP and GaAs as active material grown on InP substrates. Quaternary alloys such as InGaAsP and InGaAlAs can be grown lattice matched to InP and make it possible to change the bandgap energy and well thickness independently [31, 32]. Thick QWs up to 12 nm have been reported in InGaAsP/InGaAsP MQWs resulting in drive voltages as low as 1.2 V [31]. Similarly, InGaAlAs/InAlAs MQWs with well widths as large as 19.6 nm yielded very low-voltage EA modulators at 1.55 μm, requiring about 1 V for 10 dB on/off ratio [33]. Larger electron confinement due to increased conduction band discontinuity in this material system makes such wells comparable to narrower wells in the InGaAsP system. It is possible to obtain similar results using lattice matched InGaAs ternary QWs. However, as the wells get thicker the absorption edge shifts to longer wavelengths and operation around 1.55 μm becomes difficult. This difficulty in the ternary material can be eliminated by using tensile strained QWs [18]. A commonly used material design consists of a 0.6 μm n-InAlAs buffer, an undoped strained MQW absorption layer, a 2 μm p-InAlAs cladding layer and a p$^+$-InGaAs contact layer [34]. The strained MQW layer contains ten 8.8 nm $\mathrm{In_{0.48}Ga_{0.52}As}$ wells and 5 nm $\mathrm{In_{0.53}Al_{0.47}As}$ barriers. The wells are under 0.35% tensile strain and the barriers are under 0.5% compressive strain. The strain in the barriers

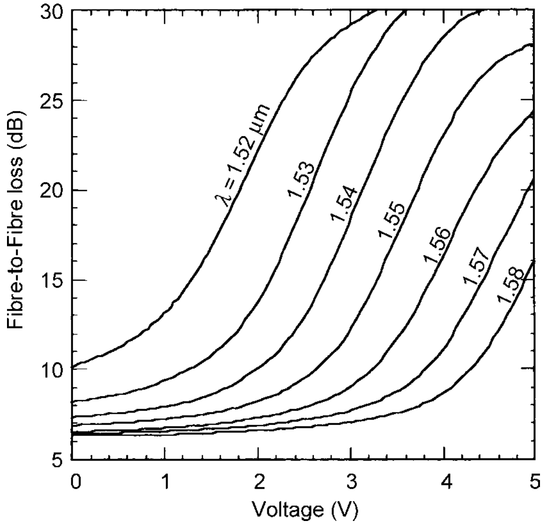


Figure B4.17. Fibre-to-fibre insertion loss of a 40 Gbit s^{-1} EA modulator for various wavelengths. The input light was TE polarized. After [34] figure 10. Reproduced by permission of IEEE.

is used for strain compensation. Proper amounts of alternating compressive and tensile strain can make the total strain in the MQW near zero, allowing large periods of strained MQWs without exceeding the critical layer thickness. Fibre-to-fibre insertion loss of a typical high-speed EA modulator as a function of external bias at different wavelengths is shown in figure B4.17 [34]. The QCSE is most pronounced for photon energies near the bandgap of the material and shows strong wavelength dependence as seen in figure B4.17. At shorter wavelengths modulation becomes more efficient, but insertion loss also increases. EA modulators are very short devices and, hence, have small device capacitance. Therefore, when driven as a lumped circuit element their speed of operation will be limited by the *RC* time constant of the circuit. Therefore, reducing the capacitance by shortening the device increases the speed of operation. Typically a 2.5 μm wide and 150 μm long device has a capacitance of about 0.33 pF, which is low enough for 20 GHz bandwidth when driven by a 50 Ω source with a 50 Ω terminating resistance across the device. Higher bandwidths are possible by using a lower terminating resistance. This reduces the *RC* time constant, hence increases the bandwidth at the expense of modulation efficiency. Electrical 3 dB bandwidths approaching 60 GHz have been reported for lumped devices [34–37]. High-speed operation requires not only a short device but also good impedance matching and proper microwave packaging. Although unpackaged devices demonstrated very high-speed operation, it was difficult to duplicate these results in packaged modules. The main source of the difficulty is cleaving and packaging devices as short as 50 μm. One solution to this difficulty is to integrate passive input and output waveguides to a very short EA section [34]. This requires etching of the MQW region outside the EA modulator section and regrowing low-loss passive waveguide sections. This technique was used to obtain overall device lengths of 1 mm, with EA modulator sections as short as 50 μm. The measured frequency response of such a 1.55 μm modulator is shown in figure B4.18, showing a 3 dB electrical bandwidth of 50 GHz. At 1.3 μm, bandwidths as large as 38 GHz have been demonstrated [38].

In unstrained material QCSE is polarization dependent. However, using strain compensated QWs polarization dependence can be significantly reduced [18, 39, 40]. The most common approach is to use about 0.5% tensile strained InGaAs wells and about 0.5% compressive strained InAlAs barriers [34, 39].

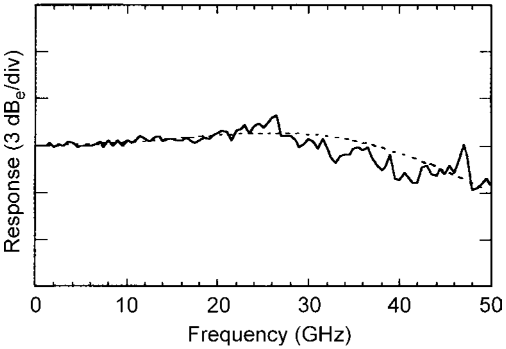


Figure B4.18. Frequency response of an EA modulator with integrated input and output optical waveguides under 1.5 V reverse bias and + 5 dBm optical input power at 1.55 μm. The dashed line shows the calculated data. After [34] figure 11. Reproduced by permission of IEEE.

Travelling wave electro-absorption modulators

One of the emerging directions in EA device research is to reduce the drive voltage and increase the bandwidth. Lumped device performance seems to be saturated around 2 V drive voltage and 50 GHz bandwidth. The main reason for that are the conflicting requirements on the length of the device. Low drive voltages require longer devices yet wide bandwidths require shorter devices. One way to address this problem is to design the device as a travelling wave modulator. As described earlier for a travelling wave device, the *RC* time constant is not the bandwidth limit and the device can be made longer without sacrificing bandwidth. However, the special structure of an EA modulator presents interesting issues. First of all, the optical propagation loss of an EA device is rather high. Therefore, increasing the length over a few hundred microns introduces excessive loss. Large device capacitance per unit length also makes it difficult to make a 50 Ω transmission line with matched velocity. Since the device is rather short, velocity matching is not an issue up to frequencies well into the sub-millimetre-wave range. The bandwidth is typically limited by the microwave loss [41]. For travelling wave electro-absorption modulator (TW-EA) electrodes, measured microwave loss coefficients of about 60–80 dB cm^{-1} at 40 GHz were reported [41]. This excessive loss is due to heavily doped layers inside the device and is another factor limiting the length of the device. Characteristic impedance values are about 25 Ω [41, 42]. However, designing the device as part of a 25 Ω transmission line makes longer devices possible and packaging easier. Recently, a 200 μm long TW-EA device with a bandwidth over 54 GHz [43] and a 300 μm long device with a bandwidth of 25 GHz [42] were reported. Both devices were polarization insensitive and operated at 1.55 μm. For 20 dB on/off ratio, drive voltages of the 200 and 300 μm devices were 3 and 1.9 V, respectively.

B4.7.2 $LiNbO_3$ modulators

Among all external modulators, $LiNbO_3$ travelling wave modulators have the most mature technology. Such modulators are commercially available from several manufacturers and are used in commercial applications. $LiNbO_3$ is a ferroelectric crystal, which is readily available commercially.

A schematic of a typical $LiNbO_3$ phase modulator is shown in figure B4.19. On the surface there are two electrodes across which the modulating voltage is applied. An optical waveguide is also formed in the crystal. The most common way of fabricating optical waveguides in $LiNbO_3$ is Ti indiffusion. Typically, Ti stripes of about 3–8 μm wide and about 0.1 μm thick are patterned on the surface of $LiNbO_3$ using liftoff. These Ti stripes are subsequently driven into $LiNbO_3$ around 1000°C for about 10 h in an oxygen

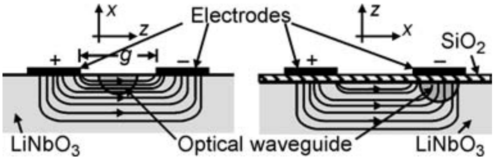


Figure B4.19. The most commonly used $LiNbO_3$ modulator configurations using (*a*) an *x*-cut, (*b*) *z*-cut crystal.

atmosphere. The resulting Ti diffusion profile translates itself into a higher index of refraction and optical guiding results. The index steps generated this way are of the order of 0.01, and typical optical propagation loss coefficients are less than $0.2\,\mathrm{dB\,cm^{-1}}$. Furthermore, the optical modes in such waveguides match rather well with the mode of a single-mode optical fibre. As a result, 5 dB fibre-to-fibre insertion is quite common for $LiNbO_3$ travelling wave modulators with 5 cm long electrodes. Other techniques of fabricating optical waveguides in $LiNbO_3$ include ion exchange, proton exchange and Ni diffusion. The effect used for modulation is the linear electro-optic effect and its specifics for $LiNbO_3$ are described next.

Electro-optic effect in LiNbO3

$LiNbO_3$ has 3*m* crystal symmetry. When an external modulating field is applied to $LiNbO_3$ its index ellipsoid is perturbed and using equations (B4.34) and (B4.35) and figure B4.15 it can be expressed as

$$(x\quad y\quad z)\begin{bmatrix} \frac{1}{n_o^2}-r_{22}\mathscr{E}_y+r_{13}\mathscr{E}_z & -r_{22}\mathscr{E}_x & r_{51}\mathscr{E}_x \\ -r_{22}\mathscr{E}_x & \frac{1}{n_o^2}+r_{22}\mathscr{E}_y+r_{13}\mathscr{E}_z & r_{51}\mathscr{E}_y \\ r_{51}\mathscr{E}_x & r_{51}\mathscr{E}_y & \frac{1}{n_e^2}+r_{33}\mathscr{E}_z \end{bmatrix}\begin{pmatrix} x\\ y\\ z\end{pmatrix}=1 \tag{B4.38}$$

where n_o and n_e are known as the ordinary and extraordinary indices of refraction, respectively. Their values are $n_e = 2.21$ and $n_o = 2.3$. The electro-optic coefficients are given as $r_{22} = 3.4\times10^{-12}\,\mathrm{m\,V^{-1}}$, $r_{13} = 8.6\times10^{-12}\,\mathrm{m\,V^{-1}}$, $r_{33} = 30.8\times10^{-12}\,\mathrm{m\,V^{-1}}$, and $r_{51} = 28\times10^{-12}\,\mathrm{m\,V^{-1}}$. In this case depending on the direction of the modulating electric field different possibilities exist. For example, if the modulating field is z directed, $\mathscr{E}_x = \mathscr{E}_y = 0$, and equation (B4.38) reduces to

$$(x\quad y\quad z)\begin{bmatrix} \frac{1}{n_o^2}+r_{13}\mathscr{E}_z & 0 & 0 \\ 0 & \frac{1}{n_o^2}+r_{13}\mathscr{E}_z & 0 \\ 0 & 0 & \frac{1}{n_e^2}+r_{33}\mathscr{E}_z \end{bmatrix}\begin{pmatrix} x\\ y\\ z\end{pmatrix}$$

$$=(x\quad y\quad z)\begin{bmatrix} \frac{1}{(n_0+\Delta n_x)^2} & 0 & 0 \\ 0 & \frac{1}{(n_0+\Delta n_y)^2} & 0 \\ 0 & 0 & \frac{1}{(n_e+\Delta n_z)^2} \end{bmatrix}\begin{pmatrix} x\\ y\\ z\end{pmatrix}=1 \tag{B4.39}$$

In this case, the index ellipsoid remains diagonal. The index of refraction for an optical wave polarized in the z direction and propagating in either the x or y direction can be found using equation (B4.39) as

$$\frac{1}{(n_e+\Delta n_z)^2}=\frac{1}{n_e^2}+r_{33}\mathscr{E}_z \quad\text{or}\quad (n_e+\Delta n_z)=\frac{n_e}{\sqrt{1+r_{33}n_e^2\mathscr{E}_z}}.$$

Since $r_{33}n_e^2E_z \ll 1$, $(n_e + \Delta n_z) \approx n_e\left(1 - \frac{1}{2}r_{33}n_e^2E_z\right)$, hence

$$\Delta n_z = -\frac{1}{2}r_{33}n_e^3E_z. \tag{B4.40}$$

Since r_{33} is the largest electro-optic coefficient, this arrangement generates the largest index change, while keeping the index ellipsoid diagonal. For this reason a z-directed modulating field, modulating a z-polarized optical wave, is the most commonly used arrangement for $LiNbO_3$ electro-optic modulators.

The two most commonly used configurations taking advantage of this arrangement are shown in figure B4.19. For x-cut crystals the optical waveguide is in the y crystal orientation. A horizontal electric field parallel to the surface of the crystal, i.e. an electric field along the z-axis, is utilized. This configuration modulates the TE polarized optical mode, in which the main electric field component of the optical waveguide mode is parallel to the surface of the crystal, i.e. it is along the z-axis of the crystal. For z-cut crystals an electric field vertical to the surface of the crystal is used. Optical waveguides are along the y-axis of the crystal as shown in figure B4.19. This field modulates the TM polarized optical mode most efficiently. For TM modes the main electric field component of the optical waveguide mode is also perpendicular to the surface of the crystal, i.e. it is along the z-axis of the crystal. For both configurations, the modulating external field and the main electric field component of the optical mode are parallel to each other and to the z-axis of the crystal. In the case of a z-cut crystal usually a low-index buffer layer, such as SiO_2, is used under the electrode. This buffer layer is used to isolate the optical mode from the metal electrode to keep the optical propagation loss low.

The presence of the applied electric field modifies the index of the material as described earlier. This in turn modifies the propagation constant of the optical mode. Since the effect is very small this perturbation can be found using a perturbation analysis. The result is [12]

$$\Delta\beta = (2\pi/\lambda)\Delta n_{\mathrm{eff}} \ \text{ with } \ \Delta n_{\mathrm{eff}} = \iint \Delta n_z |Y|^2\,\mathrm{d}S$$

where Y is the normalized electric field of the optical mode. The integration is carried out over the entire optical mode. Substituting the value of Δn_z derived in equation (B4.40) we obtain

$$\Delta n_{\mathrm{eff}} = \frac{1}{2}r_{33}n_e^3 \iint E_z |Y|^2\,\mathrm{d}S. \tag{B4.41}$$

Multiplying and dividing this equation by v/g, where v is the applied voltage and g is the electrode gap, we obtain

$$\Delta n_{\mathrm{eff}} = \frac{1}{2}r_{33}n_e^3\frac{v}{g}\left[\frac{g}{v}\iint E_z|Y|^2\,\mathrm{d}S\right] = \frac{1}{2}r_{33}n_e^3\frac{v}{g}\Gamma \tag{B4.42}$$

where Γ is known as the overlap integral and is expressed as

$$\Gamma = g\iint \frac{E_z}{v}|Y|^2\,\mathrm{d}S. \tag{B4.43}$$

Γ is proportional to the overlap of the magnitude squared normalized optical mode electric field and the normalized z-component of the applied modulating field. With this definition, we can express the K-coefficients defined in equation (B4.13) and (B4.11) as

$$K = \frac{1}{2}r_{33}n_e^3\frac{\Gamma}{g}. \tag{B4.44}$$

Therefore, the V_π of a push–pull driven $LiNbO_3$ directional coupler modulator is given as

$$V_\pi = \frac{\sqrt{3}}{2}\frac{\lambda_0}{L}\frac{g}{r_{33}n_e^3\Gamma}. \tag{B4.45}$$

Similarly that of a push–pull driven $LiNbO_3$ Mach–Zehnder modulator is

$$V_\pi = \frac{1}{2}\frac{\lambda_0}{L}\frac{g}{r_{33}n_e^3\Gamma}. \tag{B4.46}$$

Obviously in order to have a low drive voltage a large electro-optic coefficient, a large index of refraction, a large overlap integral, a long electrode length and a small electrode gap are desired. A simple back of the envelope calculation indicates that unless L is very long it is not possible to achieve low-drive-voltage modulators. However, long electrode length necessitates a travelling wave configuration as described earlier. For this reason, all practical $LiNbO_3$ electro-optic modulators are travelling wave modulators. Such modulators are described in the next section.

Travelling wave $LiNbO_3$ electro-optic modulators

A schematic of a typical $LiNbO_3$ travelling wave modulator is shown in figure B4.20 [44]. The optical structure is a Mach–Zehnder interferometer. The electrode length is of the order of centimetres, hence velocity matching is essential in such a modulator.

The dielectric constant of $LiNbO_3$ shows a large amount of dispersion going from microwave to optical frequencies due to a large ionic contribution to its dielectric constant. For *z*-cut crystals the

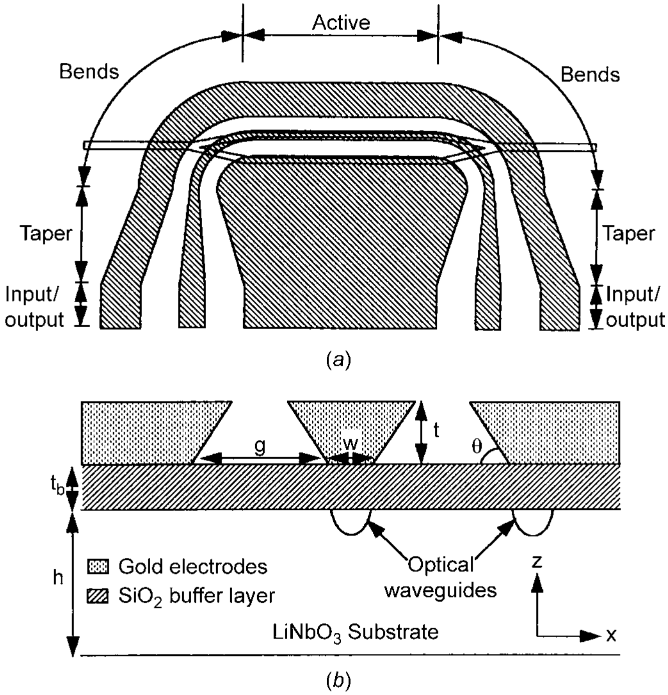


Figure B4.20. (*a*) Top view, (*b*) cross sectional schematics of a *z*-cut *y*-propagating $LiNbO_3$ travelling wave modulator. After [44] figure 1. Reproduced by permission of IEEE.

microwave relative dielectric constants of $LiNbO_3$ parallel and perpendicular to the crystal are $\varepsilon_{ry} = 43$ and $\varepsilon_{rz} = 28$, respectively. For a transmission line running along the y-direction one can define an effective relative dielectric constant under the quasi-static approximation as [45]

$$\varepsilon_{\text{reff}} = \sqrt{\varepsilon_{ry}\varepsilon_{rz}} \tag{B4.47}$$

which is around 35. The most commonly used electrode for such modulators is the coplanar transmission line (CPW). The effective microwave index of a CPW on $LiNbO_3$ with zero conductor thickness and top dielectric as air is [46]

$$n_{\mu} = \sqrt{\frac{\varepsilon_{\text{reff}} + 1}{2}} = \sqrt{\frac{\sqrt{\varepsilon_{ry}\varepsilon_{rz}} + 1}{2}}. \tag{B4.48}$$

This value is larger than 4. On the other hand, the commonly assumed effective index of an optical mode in a Ti indiffused $LiNbO_3$ optical waveguide is 2.15. As a result, an electrical signal applied to the electrode will travel significantly slower than the optical wave. Therefore, velocity matching in $LiNbO_3$ modulators requires increasing the velocity of propagation of the microwave electrode.

The most common way of achieving this is to use a SiO_2 buffer layer under the electrode and to increase the thickness of the conductors. The characteristic impedance, Z_0, and the velocity of propagation, v_{μ}, of a transmission line are given as

$$Z_0 = \sqrt{\frac{l}{C}} \quad v_{\mu} = \frac{1}{\sqrt{lC}}, \tag{B4.49}$$

where l and C are the inductance and capacitance per unit length of the transmission line, respectively. Under the quasi-static approximation, l is found using exactly the same line geometry in which the dielectrics are replaced by air. Since electro-optic materials used are nonmagnetic, l of this air line is the same as the l of the original line. But the capacitance per unit length of the airline, C_a, is different. Such a line with uniform air dielectric supports a TEM mode with velocity of propagation the same as the speed of light in air, c. Hence

$$c = \frac{1}{\sqrt{lC_a}} \quad \text{or} \quad l = \frac{1}{c^2 C_a}. \tag{B4.50}$$

Combining these equations together we find that

$$v_{\mu} = c\sqrt{\frac{C_a}{C}} \quad \text{and} \quad Z_0 = \frac{1}{c\sqrt{CC_a}}. \tag{B4.51}$$

The relative dielectric constant of SiO_2 is 3.9, which allows filling part of the transmission line with a low-index medium. This in turn reduces C hence v_{μ} and Z_0 are increased. Placing a dielectric like SiO_2 under the metal electrode also helps to reduce the overlap of the optical mode with the metal electrode as described earlier. This helps to reduce the optical propagation loss significantly especially for z-cut devices. However, thick SiO_2 layers are not desirable since part of the electrode voltage that drops across this layer reduces the electric field intensity in $LiNbO_3$ and hence the modulation efficiency of the device. Typical SiO_2 layer thickness is about 1 μm. The additional velocity increase is typically achieved by increasing the conductor thickness. Increasing the conductor thickness increases the electric field strength in the slots between the conductors. This in turn increases C_a, hence v_{μ} increases and Z_0 decreases. Typical conductor thicknesses range from 10 to 20 μm. For such thick electrodes, the slope of the sidewalls starts to affect the electrode characteristics [44]. Although it is possible to achieve velocity matching this way, the undesirable side effect is the reduction in the characteristic impedance of the electrode. It is possible to adjust the gap and the width of the centre conductor of the CPW to match v_μ and Z_0 simultaneously, but this requires rather narrow gap and width values [47]. As a result microwave electrode loss increases and becomes the limiting factor for the bandwidth. This difficulty was solved by introducing the ridge structure [48].

Figure B4.21 illustrates the basic idea behind this approach [48]. Compared to the conventional modulator, the ridge removes the high-microwave-dielectric-constant $LiNbO_3$ between the conductors. As a result, C reduces and both v_μ and Z_0 are increased simultaneously. This allows velocity matching without sacrificing impedance matching. Furthermore, a high dielectric constant in the ridge helps to confine the electric field lines under the electrode such that field becomes almost vertical under the electrode. As a result, the desired vertical component of the electric field overlaps better with the optical mode improving the efficiency of the modulator. This is especially true for the *z*-cut devices. Detailed analysis of this modulator geometry was reported in [49].

Travelling wave $LiNbO_3$ modulators with very wide bandwidths have been demonstrated. The frequency dependence of the modulation response of a modulator fabricated using the ridge structure is shown in Figure B4.22.

For this modulator $t_b = 1.0\,\mu m$, ridge height, $t_r = 4.0\,\mu m$, $t_m = 20\,\mu m$, $L = 2\,cm$, $W = 8\,\mu m$, and $G = 25\,\mu m$. The measured electrical and optical bandwidths are 75 and 110 GHz, respectively, and V_π is 5.1 V [50]. If the length of the same device is increased to 3 cm, V_π decreases to 3.5 V but the electrical and optical bandwidths also decrease to 30 and 45 GHz, respectively. Recently, a careful loss measurement up to 110 GHz revealed that up to 20 GHz loss is dominated by conductor losses [51]. Above 20 GHz

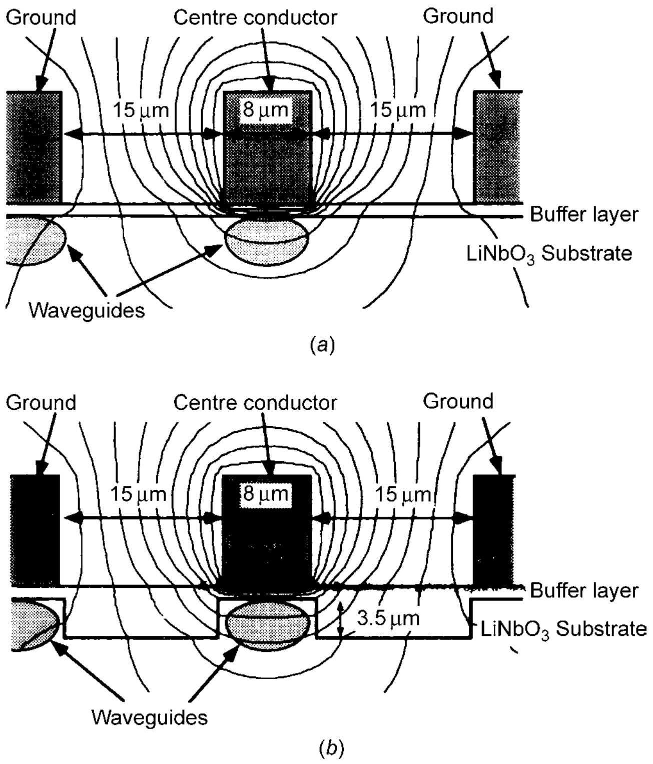


Figure B4.21. Cross sectional schematic and potential distribution around the centre conductor of (*a*) a conventional modulator and (*b*) a ridge structure. After [48] figure 3. Reproduced by permission of IEEE.

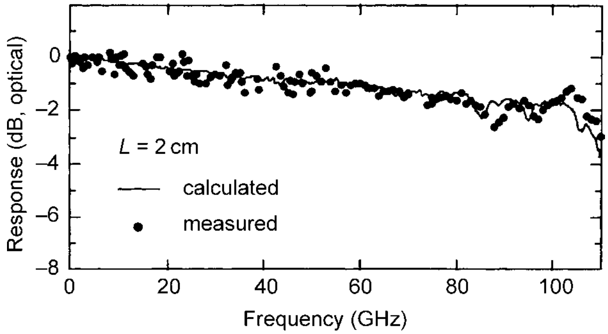


Figure B4.22. Modulation response as a function of frequency for a broadband $LiNbO_3$ travelling wave modulator employing ridge structure. After [51] figure 9. Reproduced by permission of IEEE.

dielectric and radiation losses become important. The SiO_2 buffer layer is found to have a loss tangent four times higher than that of the $LiNbO_3$ substrate [51]. Therefore, the quality of this buffer layer needs to be carefully controlled. Another loss component is the coupling to substrate modes in a CPW structure [52]. The usual technique to eliminate this loss is to thin the $LiNbO_3$ substrate. Typical substrate thickness used is less than 0.5 mm. The thinner the substrate, the higher the frequency at which this coupling occurs. Also keeping a large part of the substrate covered with metal helps to eliminate this coupling. CPW electrodes provide an advantage in this regard. In the measurements in [51], several narrowband electrode loss increases at 85, 95 and 105 GHz were reported. This observation was attributed to coupling to substrate modes in the electrical connectors.

The recent results in the bandwidth of $LiNbO_3$ modulators are impressive. However, drive voltages required to drive such modulators at high frequencies are still high. Present day electronics is expected to generate about 3.5 V drive voltages at 40 GHz and even less at higher frequencies. That means the drive voltages of existing modulators need to be reduced even further. One obvious way is to increase the electrode length. In a recent work, a reflection type travelling wave modulator was reported [53]. This approach reflects the electrical and optical signals from the cleaved and polished edge of the substrate. The reflection doubles the interaction length with the electrode. For an electrode and interaction length of 5.3 and 10.6 cm, respectively, V_π was 0.89 V at 1.3 μm.

$LiNbO_3$ is susceptible to what is known as photorefractive effect [54]. Under optical illumination, electrons from ionized impurity centres (typically Fe^{2+}) are excited to the conduction band. Then they move around until they are recaptured by traps. This creates charge separation, which creates large electric fields inside the material. These electric fields change the index through the electro-optic effect. Propagation loss also changes. This is a time dependent process depending on the charge dynamics and changes in the illumination levels. This effect is most significant at wavelengths shorter than about 870 nm [55]. It becomes less of an issue at infrared wavelengths like 1.55 μm for typical power levels used in communication applications. It is shown that photorefractive sensitivity of Ti indiffused $LiNbO_3$ modulators can be reduced significantly using high-temperature O_2 anneals [56]. Annealed modulators were operated at 1.3 μm at input optical powers as high as 400 mW up to 168 h with only a 3.5° drift in the modulator bias point. There was no measurable change in the bias point at 100 mW. It is also shown that annealing at high temperatures in N_2 and Ar ambient increases the sensitivity to the photorefractive effect significantly [56].

Bias stability of modern day $LiNbO_3$ devices is very good. They can operate at constant bias for thousands of hours under the control of automatic biasing circuitry. However, when bias voltages change suddenly, bias drifts with time constants from hours to many days may be observed. The major source of bias point drift was attributed to flow and redistribution of electrical charge under the application of applied voltages. The conductivity of SiO_2 buffer and $LiNbO_3$ are very high but finite. Usually a certain amount of low-mobility charge exists in these materials, at their interface or on their surface. These charges could be process related or can be generated with optical or thermal excitation during the operation of the device. When voltages are applied to the electrodes, charge starts to move and modifies the electric field applied to the optical guides. As a result, modulation characteristics change. This behaviour can be modelled using an electrical network approach [57]. Each layer in the device can be modelled by an equivalent resistance representing the flow of charge through it and each interface can be modelled by a capacitance representing the potential for charge storage. Such models are able to describe experimental observations successfully [57]. The usual method of preventing such drift involves controlling the resistivity of the buffer and matching it to the resistivity of waveguide layers and modifying the $LiNbO_3$ surface before depositing the buffer layer.

The electro-optic effect in $LiNbO_3$ depends on the polarization of the light and high-speed modulators are polarization dependent. However, polarization independent $LiNbO_3$ has been demonstrated. This usually takes advantage of the anisotropic electro-optic tensor of $LiNbO_3$. One approach relies on TE/TM mode coupling via off-diagonal elements of the electro-optic tensor [58]. In another approach, a $\Delta\beta$ coupler is used [59]. There have been other novel approaches using crossing waveguides and specially designed directional couplers [60, 61]. Digital optical switches can also be used as polarization independent modulators [62]. Polarization independent modulators usually require special crystal cuts and electrode geometries. As a result, the electro-optic coefficient used is not the largest and the electrode is not suitable for high-speed operation. This makes the efficiency and speed of such devices low.

B4.7.3 III–V compound semiconductor electro-optic modulators

Electro-optic effect in III–V compound semiconductors

III–V compound semiconductors such as GaAs, InP and their alloys have excellent optical properties due to their direct and tunable bandgap and are materials of choice in many optoelectronic components such as lasers and detectors. They also lack inversion symmetry and possess an electro-optic coefficient. It is also possible to utilize other electro-optic effects described earlier if the wavelength of operation is close to the material bandgap. As a result they are also used in modulator applications. The most commonly used optical structure is a Mach–Zehnder interferometer. There are many different ways of making optical waveguides in III–V compound semiconductors. It is possible to adjust the index of refraction of these materials by controlling the composition of their alloys. For example increasing Al composition x in an $Al_xGa_{1-x}As$ compound semiconductor decreases its index of refraction. Furthermore, $Al_xGa_{1-x}As$ is lattice matched to GaAs for all x values. By growing such layers epitaxially using techniques such as molecular beam epitaxy (MBE) or metal organic chemical vapour deposition (MOCVD), it is possible to sandwich a higher-index material between two lower-index materials. This forms a slab waveguide and provides waveguiding in the vertical direction. The most common approach to provide lateral waveguiding is to etch a rib and form what is known as a rib waveguide. The effective index under the rib is higher than the effective index outside the rib in the etched regions. This provides a lateral index step and a two-dimensional waveguide is formed. There are many other ways to provide waveguiding which include proton implantation, buried heterostructures, pn junctions and disordering. However, rib waveguides are almost the universal choice since vertical and lateral index profiles and dimensions can be independently and precisely controlled. Typical rib widths are in the 2–4 μm range and rib heights are typically less than 1 μm.

III–V compound semiconductors have $\bar{4}3m$ or zinc blende crystal structure. When an external modulating field is applied, the index ellipsoid is perturbed and using equations (B4.34) and (B4.35) and figure B4.15 it can be expressed as

$$(x \quad y \quad z)\begin{bmatrix} \frac{1}{n^2} & r_{41}\mathscr{E}_z & r_{41}\mathscr{E}_y \\ r_{41}\mathscr{E}_z & \frac{1}{n^2} & r_{41}\mathscr{E}_x \\ r_{41}\mathscr{E}_y & r_{41}\mathscr{E}_x & \frac{1}{n^2} \end{bmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 1. \tag{B4.52}$$

The refractive index value for compound semiconductors is around 3.5 at 1.55 μm. This value can be adjusted over a wide range by changing the composition of the material. The electro-optic coefficient is given as $r_{41} = 1.4 \times 10^{-12}\,\mathrm{m\,V^{-1}}$ [63]. In this case, depending on the direction of the modulating electric field different possibilities exist. For example, if the modulating field is z directed, $\mathscr{E}_x = \mathscr{E}_y = 0$, and equation (B4.52) reduces to

$$(x \quad y \quad z)\begin{bmatrix} \frac{1}{n^2} & r_{41}\mathscr{E}_z & 0 \\ r_{41}\mathscr{E}_z & \frac{1}{n^2} & 0 \\ 0 & 0 & \frac{1}{n^2} \end{bmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 1 \tag{B4.53}$$

In this case, the index ellipsoid is no longer diagonal. It can be diagonalized using a simple coordinate transformation in the xy-plane as shown in figure B4.23. The new $x'y'$-axes are simply rotated 45° with respect to the xy-axes. In the new $x'y'z$ coordinate system index ellipsoid can be written as

$$(x' \quad y' \quad z)\begin{bmatrix} \frac{1}{n^2} - r_{41}\mathscr{E}_z & 0 & 0 \\ 0 & \frac{1}{n^2} + r_{41}\mathscr{E}_z & 0 \\ 0 & 0 & \frac{1}{n^2} \end{bmatrix}\begin{pmatrix} x' \\ y' \\ z \end{pmatrix}$$

$$= (x' \quad y' \quad z)\begin{bmatrix} \frac{1}{(n+\Delta n_{x'})^2} & 0 & 0 \\ 0 & \frac{1}{(n+\Delta n_{y'})^2} & 0 \\ 0 & 0 & \frac{1}{n^2} \end{bmatrix}\begin{pmatrix} x' \\ y' \\ z \end{pmatrix} = 1. \tag{B4.54}$$

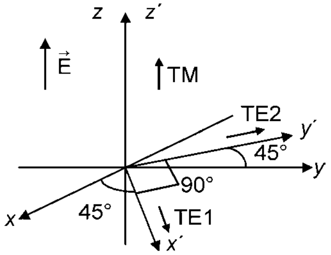


Figure B4.23. The axes of the index ellipsoid for an III–V compound semiconductor in the case of a z-directed external electric field. The dielectric tensor is diagonal in x, y, z and x', y', z' coordinate systems in the absence and the presence of the electric field, respectively. TE1, TE2, and TM show the direction of the main optical field of the two orthogonal TE and TM modes, respectively.

In this case, there is no index change for an optical wave polarized in the z-direction. The index of refraction for an optical wave polarized in the x'-direction and propagating in the y'-direction can be found using equation (B4.54) as

$$\frac{1}{(n+\Delta n_{x'})^2} = \frac{1}{n^2} - r_{41}\mathscr{E}_z \quad \text{or} \quad (n+\Delta n_{x'}) = \frac{n}{\sqrt{1-r_{41}n^2\mathscr{E}_z}}.$$

Since $r_{41}n^2\mathscr{E}_z \ll 1$, $(n+\Delta n_{x'}) \approx n\left(1+\frac{1}{2}r_{41}n^2\mathscr{E}_z\right)$, hence

$$\Delta n_{x'} = \frac{1}{2}r_{41}n^3\mathscr{E}_z. \tag{B4.55}$$

Similarly the index of refraction for an optical wave polarized in the y'-direction and propagating in the x'-direction can be found as

$$\Delta n_{y'} = -\frac{1}{2}r_{41}n^3\mathscr{E}_z. \tag{B4.56}$$

In practice, the z-axis is along the 001 crystal plane. Then the x'- and y'-axes are along 110 and $1\bar{1}0$ crystal axes. Therefore, in such material a vertically applied electric field in the 001 direction increases the index of refraction by $\Delta n_{x'}$ in the 110 direction and decreases it by $\Delta n_{y'}$ in the $1\bar{1}0$ direction. In other words, the index increases along one of the two mutually orthogonal directions parallel to the surface of the crystal. It decreases by the same amount in the other orthogonal direction. These directions correspond to the cleavage planes of the 001 oriented material. No index change is observed in the 001 direction, which implies that a vertically applied electric field to a 001 oriented crystal will only modulate the TE mode of the optical waveguide in which the main electric field component of the optical mode is either in 110 or $1\bar{1}0$ directions, in other words it is tangential to the surface. No modulation will result for the TM mode, which has its main electric field component in the 001 direction, i.e. normal to the surface of the crystal. The electro-optic coefficient in compound semiconductors is about 20 times less than that of $LiNbO_3$; however, the net index change for a given electric field is only about five times less due to the higher index of refraction of the semiconductor. It is possible to enhance this index modulation using other physical effects described earlier.

Lumped III–V compound semiconductor modulators

Electric fields that exist in the depletion regions of either pn junctions or Schottky barriers are used to modify the index of compound semiconductors. Figure B4.24 shows a phase modulator using the electric field in the reverse biased pn junction of a GaAs/AlGaAs bulk heterostructure.

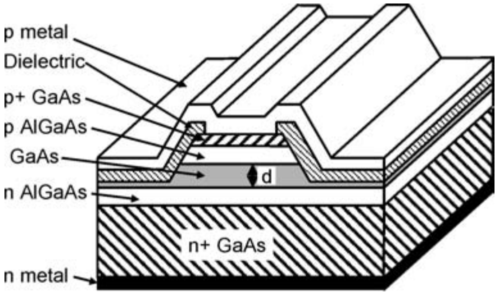


Figure B4.24. Schematic of a III–V compound semiconductor waveguide phase modulator.

The high-index GaAs layer of thickness *d* in between two lower-index AlGaAs layers provides vertical waveguiding. For lateral confinement etching the layers outside the rib of width *w* forms a rib waveguide. Ohmic contacts are formed to the top p^+ GaAs contact layer and to the backside of the n^+ substrate. A dielectric isolation layer, such as SiO_2, covers the sidewalls and the etched regions. By reverse biasing the pn junction a large electric field is generated in the depletion region. Typically the core of the waveguide, in this case the GaAs layer, is undoped. Hence, a fairly constant electric field is generated across the core of the waveguide. This electric field generates an index change through the electro-optic effects described earlier. These are the linear electro-optic effect, electrorefractive effect, plasma effect, and band filling effects. QWs in place of bulk material are also used to enhance the electrorefractive effect. We can express the resulting effective index change using equations (B4.29), (B4.42) and (B4.55) as

$$\Delta n_{\mathrm{eff}} = \left(\frac{1}{2} r_{41} n^3 \frac{v}{g} \Gamma + G(\lambda)\left(\frac{v}{g}\Gamma\right)^2\right). \tag{B4.57}$$

In equation (B4.57) it is assumed that the linear electro-optic and electrorefractive effects are the dominant effects. The basic phase modulator shown in figure B4.24 can easily be extended into an amplitude modulator using either a Mach–Zehnder or directional coupler design. The presence of doped layers may increase the optical insertion loss. One of the doped layers can be eliminated by replacing one side of the junction with a Schottky barrier. As a matter of fact totally undoped material can be used if Schottky barriers replace both doped layers. One such modulator was fabricated using substrate removal [64]. The top and cross sectional schematic of a substrate-removed modulator is shown in figure B4.25.

In this case an undoped GaAs/AlGaAs epilayer grown on a GaAs substrate is etched to form rib waveguides. An AlAs etch stop layer is also grown between the substrate and the epilayer. Then an Schottky barrier is formed on top of the rib. Next the entire wafer is glued onto a transfer substrate using a polymer bonding layer. In this case the polymer is benzocyclobutane (BCB). After this bonding the growth substrate is etched away. The substrate etch stops on the AlAs etch stop layer. Next the etch stop layer is also etched away. This exposes the backside of the epilayer. Another Schottky electrode is formed on the backside. Hence in between two electrodes on either side of the epilayer two back-to-back Schottky diodes are formed. A voltage applied to the electrodes reverse biases one of the Schottky diodes and the electric field formed in its depletion region generates index changes. The thin undoped

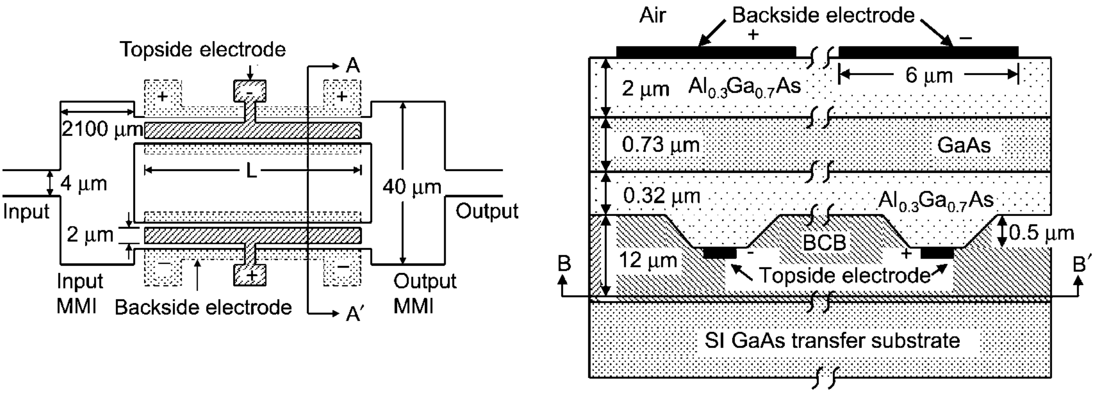


Figure B4.25. The top and cross sectional schematic of a substrate removed GaAs/AlGaAs electro-optic modulator. After [64] figures 1 and 2. Reproduced by permission of IEEE.

GaAs/AlGaAs layers self-deplete due to Fermi level pinning at the surfaces and behave effectively as a dielectric material with electro-optic properties. One big advantage of such a design is the ability to bias both arms of the interferometer independently. As a result true push–pull operation is possible. This creates a zero-chirp modulator. Using appropriate bias on both arms any amount of chirp can be obtained. The modulator reported in [64] had V_π of 8.7 and 17.8 V cm when driven push–pull or single arm, respectively.

Travelling wave III–V compound semiconductor modulators

For high-speed operation travelling wave designs are used. The index of the bulk compound semiconductors is isotropic. Furthermore, the refractive index variation between microwave and optical frequencies is rather small. The optical index of refraction is around 3.4 and the relative dielectric constant at microwave frequencies is about 13, which corresponds to an index of 3.6. Since it is a lot easier to form electrodes on the surface of the crystal the most commonly used electrode is either a CPW or a CPS structure. The optical signal is entirely confined in the semiconductor, which has a refractive index of about 3.4. On the other hand, the microwave and millimetre wave electric field fringes into the air and experiences an effective index between that of the air and semiconductor. For example, for a coplanar line of zero conductor thickness the effective dielectric constant is the arithmetic mean of the dielectric constants of the air and the semiconductor as indicated earlier in equation (B4.48). Therefore, the effective dielectric constant is around 7, which corresponds to a microwave index of about 2.65. Therefore, there is about 38% of index mismatch between the optical and microwave signals. This requires about 23% velocity reduction. Therefore, in III–V compound semiconductors velocity matching requires slowing down of the microwave signal. The most commonly used technique to slow the microwave signal is to use a slow wave transmission line [10, 65–67]. Such lines are obtained by periodically loading a uniform transmission line. The loading element is typically a capacitor. This can be achieved using either doped or undoped epitaxial layers.

Figure B4.26 shows the schematic of a travelling wave Mach–Zehnder modulator using doped layers [65]. The optical structure is a Mach–Zehnder interferometer utilizing multimode interference sections at the input and output for power splitting and combining. A GaAs/AlGaAs epitaxial layer is grown on a semi-insulating (SI) GaAs substrate. Underneath the GaAs core there is a buried n^+ layer which acts as a ground plane. The main electrode is a coplanar strip line. This electrode is periodically loaded by narrow and small capacitive elements. In [65], the capacitive elements used are the reverse biased capacitance of a Schottky–i–n junction as shown in figure B4.26. The advantage of this approach is to utilize the large vertical electric field existing in the reverse biased Schottky–i–n junction. This field overlaps very well with the optical mode enabling low drive voltages. Furthermore, these capacitive elements do not carry any of the axial currents in the transmission line. Hence increase in the microwave loss of the electrode due to loading is marginal.

Such devices with a total of 1 cm electrode length and loading segment lengths of 0.5, 0.6 and 0.7 mm were fabricated and characterized [65]. DC V_π values at 1.15 μm for 0.5, 0.6 and 0.7 mm segment-length electrodes were 5.7, 4.25 and 4.24 V, respectively. Corresponding bandwidths were >26.5, 25.0 and 22.5 GHz, respectively. Recently, such a device demonstrated an electrical 3 dB bandwidth of 50 GHz and V_π of 13 V at 1530 nm [66].

Another approach for GaAs travelling wave modulator design is to use unintentionally doped epitaxial layers [10, 68]. Such GaAs/AlGaAs layers self-deplete due to Fermi level pinning at the surface and the depletion originating at the semi-insulating substrate interface and behave very similarly to low-loss dielectric materials. As a result, optical and microwave losses become very low. The required velocity slowing can be achieved using a properly designed electrode. A schematic of such a device is shown in figure B4.27. The optical structure is a Mach–Zehnder interferometer. A (001) oriented

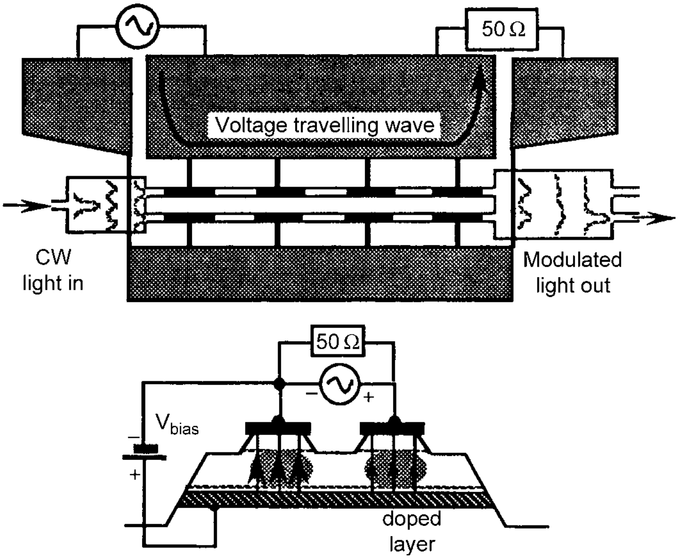

Voltage travelling wave
50 Ω
CW light in
Modulated light out
50 Ω
V_{bias}
doped layer

Figure B4.26. Schematic of travelling wave Mach–Zehnder modulator using capacitively loaded coplanar strip line. After [66] figure 1. Reproduced by permission of IEEE.

GaAs/AlGaAs heterostructure grown by MBE on semi-insulating GaAs provides vertical optical waveguiding. The lateral waveguiding is obtained by etching ridges down into the top AlGaAs layer. The microwave electrodes form Schottky contacts with the epitaxial layers. A voltage applied between the electrodes biases two back-to-back Schottky diodes. The conductivity of self-depleted epilayers is so low that, for frequencies larger than 1 MHz, the epilayers start to behave like slightly lossy dielectrics. Hence, the situation becomes identical to electrodes on an insulating dielectric. This makes it possible to

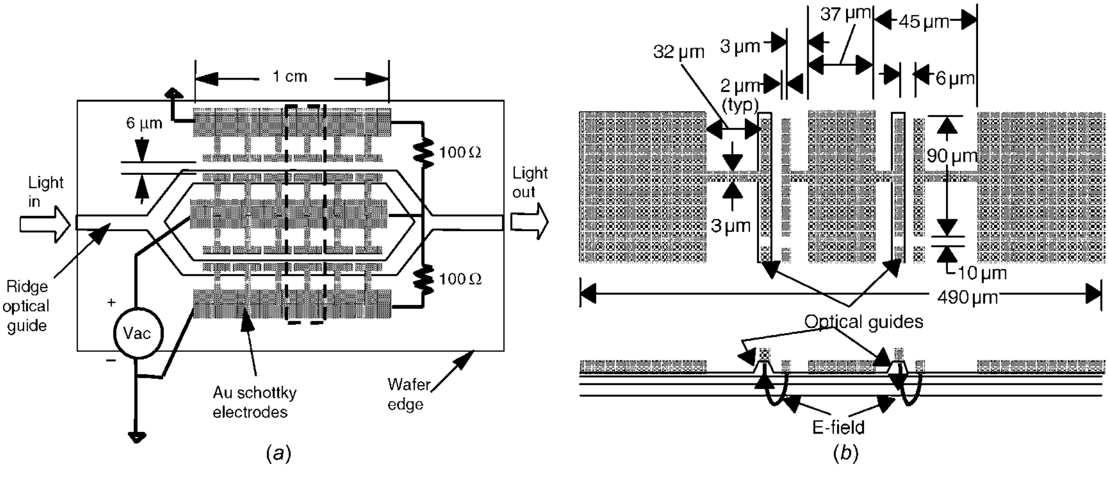

1 cm
6 μm
100 Ω
100 Ω
Light in
Light out
Ridge optical guide
Vac
Au schottky electrodes
Wafer edge
(a)
37 μm
45 μm
32 μm
3 μm
2 μm (typ)
6 μm
90 μm
3 μm
10 μm
490 μm
Optical guides
E-field
(b)

Figure B4.27. Schematic top view of a GaAs/AlGaAs modulator using unintentionally doped layers, (*b*) top view schematic of the modulator section delineated by the dashed line in (*a*) together with a cross sectional schematic. After [68] figures 1 and 2. Reproduced by permission of IEEE.

apply mainly (001) directed electric fields of opposite polarity on the optical guides as shown in figure B4.27. This generates phase shifts of opposite sign on both arms through the linear electro-optic effect, creating a net differential phase shift between the arms of the interferometer. Hence, push–pull modulation results. The device electrode is a modified coplanar line, in which T-rails stem from either side of the centre conductor and from the inner side of the both ground planes [69]. These T-rails form tiny capacitors, which periodically load the line, increase its capacitance per unit length, and, thus, slow the microwave signal propagating on the electrode. For this structure, axial transmission line currents cannot flow along these T-rails. Only displacement current flows in these tiny capacitors. Therefore, the distance between the centre conductor and the ground plane of the unloaded line determines current crowding and the microwave loss. One can make that gap large to keep microwave loss low [69]. On the other hand, the gap between the T-rails determines the field applied to the optical guides for a given voltage. In this design this gap can be reduced significantly with a very small increase in the microwave loss [69]. As a result the electrode gap has been decoupled from the electrode loss. The small-signal electrical bandwidth of this device at 1.55 μm is in excess of 40 GHz.

Polarization independent operation of III–V electro-optic modulators can be achieved either designing the material using an appropriate amount of strain, or using novel device ideas. Strain may provide equal electro-optic coefficients for TE and TM modes making the operation the same for two orthogonal polarizations. In bulk material it is possible to get TE/TM coupling for certain applied field orientations. In one approach polarization dependence was obtained using near degenerate TE and TM modes and controlling the coupling between them [70]. In this case electrodes were suitable for travelling wave operation [71] and such a device has potential for broadband polarization independent operation.

B4.7.4 Polymer modulators

Organic polymers have many attractive features for integrated optical applications [72]. It is possible to form multilayer polymer stacks by spin deposition and curing on a large variety of substrates or even on fabricated circuits [73]. They can be patterned using several different techniques such as photo-bleaching and reactive ion etching. They present good optical properties, such as low propagation loss and low index of refraction very close to that of the single-mode fibre. They also have low dielectric constant, which is important for high-speed devices. These properties resulted in passive low-loss polymer optical waveguides that can couple to single-mode optical fibres very efficiently [74, 75].

Polymers can also be made electro-optic and used in active devices. Certain molecules known as chromophores possess large ground state electric dipole moments and exhibit large optical nonlinearities. This microscopic nonlinearity can be converted to a macroscopic nonlinearity by mixing them with polymers and aligning their dipole moments. The most common method is to take a chromophoric polymer film and apply a strong electric field across it while keeping the film heated at a high temperature near its glass transition temperature. At such high temperatures randomly aligned individual chromophores are able to move in the polymer and align themselves with the externally applied electric field due to their dipole moments. After this alignment if the film is cooled to room temperature while applied electric field is present ordering of chromophores is achieved. This creates a macroscopic nonlinearity that can be used for electro-optic modulation. This process is known as high-temperature poling. Poling temperatures and fields depend on the particular film, but are typically about 100–200°C and 100–200 V μm^{-1}.

Most common techniques of poling a polymer film are either corona or electrode poling. In corona poling, corona discharge is used to create large poling fields over large areas. In electrode poling one approach is to spin coat the lower cladding and core on a ground electrode. Then poling electrodes are formed on top of the cladding. After poling these electrodes are removed and the upper cladding is spin coated. Alternatively, whole stack can be spin coated, followed by poling electrode formation, poling and removal of poling electrodes. The conductivity of upper and lower cladding regions of the polymer waveguide should be larger than the core of the waveguide for most of the poling voltage to drop across the electro-optic core and result in effective poling. Usually, the electro-optic coefficient increases linearly with poling field but so does the optical loss and birefringence.

After poling electro-optic activity is observed. For modulating fields applied in the poling field direction, two different electro-optic coefficients are observed depending on the polarization of the optical mode. The coefficient for TM polarization, i.e. when the optical field is polarized in the modulating field direction, is r_{33}. The coefficient for TE polarization, i.e. when the optical field is polarized perpendicular to the modulating field direction, is r_{13}. Typically for polymers r_{33} is about three times larger than r_{13}. At the present time, for most of the reported modulators, r_{33} values range between 1 and 20 pm V^{-1} although values as high as 67 pm V^{-1} have been reported [76]. The index change for an applied electric field is given by equation (B4.42). Polymer indices of refraction are around 1.6, which makes a polymer with $r_{33} \cong 12$ pm V^{-1} equivalent to bulk GaAs as far as index change is concerned. To take advantage of high r_{33}, modulating field should be applied in the poling field direction and the optical mode should be polarized in the same direction. For these considerations the most commonly used electrode is microstrip and the optical polarization is TM.

Figure B4.28 shows a travelling wave polymer modulator [79]. The optical structure is the Mach–Zehnder interferometer. The optical waveguides were fabricated by spin coating three layers of polymers that act as claddings and the core of the optical waveguide onto a high-resistivity Si wafer which was coated with a patterned gold plated film. Exposing the polymer film to a high-intensity light at the appropriate wavelength reduces the refractive index in the exposed areas, which is known as photo-bleaching. After poling, photo-bleaching was used to form channel optical waveguides. The thickness of the polymer stack was 6.5 μm. Tapered coplanar lines were used to couple in and out of the 12 mm long microstrip electrode. V_π of this modulator was 10 V. Low dielectric constant and low dielectric constant dispersion of polymers from microwave to optical frequencies offer an advantage in the design of travelling wave modulators. It is possible to get very good velocity matching using a microstrip electrode. For example, using a commonly quoted value of 1.6 for the optical index and 2.9 for the microwave relative dielectric constant and a 10 μm thick polymer stack, a 50 Ω microstrip line would have about 25 μm strip width and less than 5% velocity mismatch. In the case of the modulator shown in figure B4.28 index mismatch was estimated to be 0.03. The bandwidth is wider than 40 GHz. Other polymer modulators with

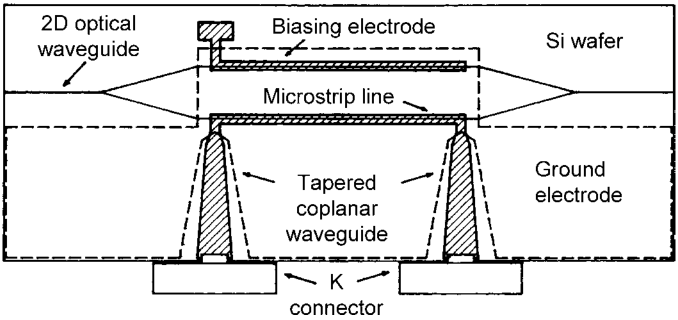


Figure B4.28. Schematic view of a polymeric travelling wave modulator. After [79] figure 1. Reproduced by permission of AIP.

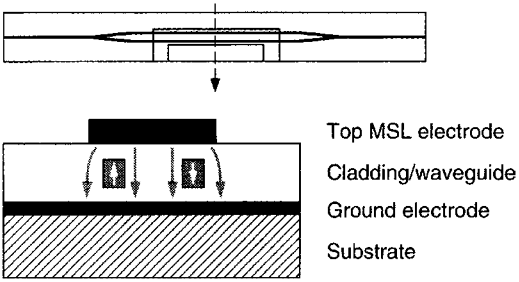


Figure B4.29. Top and cross sectional view of the device showing the optical push–pull operation in a polymer electro-optic Mach–Zehnder modulator and a single microstrip line driving circuit. The white and grey arrows represent poling and modulation field directions, respectively. After [81] figure 1. Reproduced by permission of IEEE.

bandwidths wider than 40 GHz have been reported [77, 78]. Furthermore, polymer modulators operating in the W band (75–110 GHz) have been reported demonstrating intrinsic high-frequency response of electro-optic polymers [77, 78].

If corona poling or a single poling electrode is used, chromophores in both arms of the Mach–Zehnder are aligned in the same direction. This makes it difficult to make a push–pull modulator using a single microstrip electrode and usually only one arm of the interferometer is phase modulated [79, 80]. It is possible to use two electrodes driven 180° out of phase on both arms of the interferometer to get push–pull operation and halve the drive voltage. This requires broadband 3 dB microwave hybrid and makes only half the power available to each arm without considering the loss of the hybrid. Another way of obtaining push–pull operation is to pole the arms of the Mach–Zehnder in opposite directions as shown in Figure B4.29 [81]. In this case poling electrodes had to be closely spaced in order to form a 50 Ω microstrip line. The polymer stack is about 10 μm thick, which makes the required microstrip width about 30 μm, and implies that the separation between the arms of the interferometer should be less than that. Therefore, the poling electrodes should be less than 30 μm apart. Since voltages approaching 1000 V must be applied to create poling fields in the range of 100 V μm^{-1} across the 10 μm film, the air breakdown between closely spaced poling electrodes becomes an issue. In this case, this difficulty was addressed by using a fused silica piece on top of the poling electrodes to modify the field distribution in the gap and the sample was placed in an inert gas (SF_6) ambient to increase the breakdown field. Modulators with 2 cm long microstrip electrodes had V_π values of 40 and 20 V for one-arm and push–pull operation, respectively, at 1.3 μm [81]. This corresponds to an r_{33} value of about 7 pm V^{-1}. The typical loss of a pigtailed modulator was about 12 dB. Loss of 6 dB was attributed to the on-chip optical loss of the 3 cm long interferometer and the rest was due to mode mismatch between the fibre and the optical waveguide. The additional processing associated with poling electrode formation can be eliminated if device electrodes can be used as poling electrodes. By applying 180° out of phase pulsed voltages to the electrodes, it is possible to pole both arms of the interferometer sequentially [82]. In this way, the maximum voltage between the electrodes is reduced from $2V_p$ in the case of DC poling to V_p in the case of pulse poling, where V_p is the poling voltage applied across one arm of the interferometer. The field strength between the device electrodes during poling can be further reduced by increasing the separation between electrodes. One approach uses two parallel microstrip lines of 100 Ω characteristic impedance each [83]. This way it was possible to get a separation of 200 μm between the arms of the interferometer.

Chromophores used in electro-optic polymers have certain optical transitions in the wavelength range used in optical applications. For example, it is reported that Disperse Red 19 (DR19) chromophores have a peak absorption wavelength around 470 nm corresponding to $\pi-\pi^*$ transition [84]. Therefore any optical energy present around this wavelength excites the chromophore which loses its orientation when it relaxes. This results in photoinduced relaxation, which reduces the electro-optic coefficient, absorption and the optical index. This effect is shown to be polarization dependent due to large optical dipole moment associated with the chromophore. At wavelengths away from the absorption peak this effect decreases significantly. However, optical nonlinearity of the material results in second-harmonic generation, which in turn generates optical power within the absorption band even for near infrared excitations [85]. Typically, second-harmonic generation efficiency in commonly used optical waveguides is very low due to poor phase matching. Nevertheless, this places an upper limit on the maximum power handling capacity of the optical waveguide. For example, at 1320 nm for 10 mW in a 10 μm^2 cross section waveguide, device lifetime is estimated to be larger than 10^4 h for a 30% decay in the electro-optic efficiency [84]. On the other hand, lifetime at 1064 nm was estimated to be ~1 h due to the 532 nm second harmonic [84]. Recently double-end cross-linked polymers were found to be stable when exposed to 250 mW at 1.32 μm for 1 week [86]. They were also found to be thermally stable when baked at 100°C in air for more than 2000 h. r_{33} of these materials are quoted to be 6 pm V^{-1}.

Long-term bias stability is another very important concern. Polymer layers have finite conductivity which affects electrical response under DC and AC applied biases. This effect was studied using an equivalent network approach [57, 87, 88]. Every layer of the optical waveguide can be modelled as a parallel *RC* circuit. The resistance and capacitance of each layer depend on its conductivity, σ, and dielectric constant, ϵ, respectively. At low frequencies and low DC biases, the fraction of the applied voltage across the electro-optically active core layer depends on the relative conductivity of the layers. For low-bias operation, conductivity of the electro-optic layer is desired to be much smaller than other layers. This condition is also required for effective poling. However, it is shown that this would generate a bias drift that depends on the amplitude of the AC voltage swing. This could be a problem in practical operation requiring continuous bias point monitoring and adjustment. This effect has a ω^{-1} dependence and can be made zero by matching the relaxation constants (σ/ε) for all layers [87].

Today, the electro-optic polymer topic is a very intense research area. One direction is to improve material stability and electro-optic properties of polymers. Another direction seems to be the elimination of high-temperature electric field poling. This would enable room-temperature processing, but requires electro-optically active, or ordered, material to start with. This can be achieved by depositing the polymer film monolayer by using Langmuir–Blodgett films and relying on self-ordering of the chromophores in each monolayer.

B4.8 Summary

Optical modulators are key components for fibre optic networks. This chapter has described the basic modulator specifications; phase and different types of amplitude modulators, as well as physical effects used in modulators. In addition, examples from different competing modulator technologies are given. Of these electro-absorption modulators offer compact size, low drive voltage and high bandwidth. However, modulation characteristics depend critically on wavelength and temperature. Furthermore, material preparation requires critical layer thickness and composition control. Lumped EA modulator modules with electrical bandwidths approaching 60 GHz and drive voltages around 2 V for 20 dB extinction have been demonstrated at 1.55 μm. Their fibre-to-fibre insertion loss is around 10 dB and polarization independent operation has been achieved. Drive voltage and bandwidth can be improved further using the TW-EA modulator approach. $LiNbO_3$ has the most mature technology. $LiNbO_3$ travelling wave modulators are commonly used in industrial applications. They have higher drive voltages, which requires a modulator driver. Such modulators with electrical bandwidths as high as

75 GHz and V_π around 5 V have been demonstrated. Their fibre-to-fibre insertion loss is around 5 dB. Further V_π reduction is being pursued. III–V compound semiconductor electro-optic modulators offer excellent bandwidth at the expense of higher drive voltages and insertion loss. Electrical bandwidths up to 50 GHz have been realized. V_π values are around 15 V and fibre-to-fibre insertion loss is in the 10–15 dB range. Efforts to reduce the drive voltage and fibre-to-fibre insertion using novel processing techniques are under way. Polymers offer the promise of low-cost technology and presently are an active research area. High-speed capability of electro-optic polymers has been demonstrated. Polymer modulators with electrical bandwidths exceeding 40 GHz has been reported. V_π values are around 10 V and fibre-to-fibre insertion loss is about 10 dB. In all technologies drive voltage reduction while improving the speed of operation remains as a challenge.

References

[1] Dagli N 1999 Wide bandwidth lasers and modulators for RF photonics *IEEE Trans. Microwave Theory Technol.* **47** 1151–1171

[2] Cox C III, Ackerman E, Helkey R and Betts G 1997 Techniques and performance of intensity modulation direct detection analog optical links *IEEE Trans. Microwave Theory Technol.* **45** 1375–1383

[3] Blumenthal D J, Carena A, Curri V and Humphries S 1999 All-optical label swapping with wavelength conversion for WDM-IP networks with subcarrier multiplexed addressing *IEEE Photon. Technol. Lett.* **11** 1497–1499

[4] Zmuda H and Toughlian E N (ed) 1994 *Photonic Aspects of Modern Radar* (Boston: Artech)

[5] Agrawal G P 1992 *Fibre Optic Communication Systems* (New York: Wiley–Interscience)

[6] Haus H A 1984 *Waves and Fields in Optoelectronics* (Englewood Cliffs: Prentice-Hall) chapter 7

[7] Koyama F and Iga K 1988 Frequency chirping in external modulators *J. Lightwave Technol.* 1 **6** 87–93

[8] Kubota K, Noda J and Mikami O 1980 Traveling wave optical modulator using a directional coupler $LiNbO_3$ waveguide *IEEE J. Quantum Electron.* **16** 754–760

[9] Rigrod W W and Kaminow I P 1963 Wide-band microwave light modulation *Proc. IEEE* **51** 137–140

[10] Spickermann R, Sakamoto S R and Dagli N 1997 GaAs/AlGaAs traveling wave electrooptic modulators *Optoelectronic Integrated Circuits Conference, SPIE International Symposium on Optoelectronics'97* (San Jose, CA, Feb. 8–14, 1997) paper 33

[11] Spickermann R, Sakamoto S R and Dagli N 1996 In traveling wave modulators which velocity to match? *IEEE/LEOS 1996 Annual Meeting* (Boston, MA, Nov. 18–21) paper WM3

[12] Coldren L A and Corzine S W 1995 *Diode Lasers and Photonic Integrated Circuits* (New York: Wiley–Interscience)

[13] Franz W 1958 *Z. Naturf.* **13a** 484

[14] Keldysh L V 1958 *Sov. Phys.–JETP* **7**

[15] Anselm A 1981 *Introduction to the Semiconductor Theory* (Englewood Cliffs: Prentice-Hall) p 447

[16] Wang S 1989 *Fundamentals of Semiconductor Theory and Device Physics* (Englewood Cliffs: Prentice-Hall) p 619

[17] Shinado M and Sugano S 1966 *J. Phys. Soc. Japan* **21** 1936

[18] Ido T, Sano H, Tanaka S, Moss D J and Inoue H 1996 Performance of strained InGaAs/InAlAs multiple quantum well electroabsorption modulators *J. Lightwave Technol.* **14** 2324–2331

[19] Bastard G, Mendez E E, Chang L L and Esaki L 1983 Variational calculations on a quantum well in an electric field *Phys. Rev. B* **28** 3241–3245

[20] Pankove J I 1971 *Optical Processes in Semiconductors* (New York: Dover) pp 74–75

[21] Hunsperger R G 1982 *Integrated Optics: Theory and Technology* (Berlin: Springer) pp 75–79

[22] Alping A and Coldren L A 1987 Electrorefraction in GaAs and InGaAsP and its application to phase modulators *J. Appl. Phys.* **61** 2430–2433

[23] Mendoza-Alvarez J G, Coldren L A, Alping A, Yan R H, Hausken T, Lee K and Pedrotti K 1988 Analysis of depletion edge translation lightwave modulators *J. Lightwave Technol.* **6** 793–808

[24] Zucker J E, Bar-Joseph I, Miller B I, Koren U and Chemla D S 1989 Quaternary quantum wells for electro-optic intensity and phase modulation at 1.3 and 1.55 μm *Appl. Phys. Lett.* **54** 10–12

[25] Mendoza-Alvarez J G, Yan R H and Coldren L A 1987 Contribution of the band filling effect to the effective refractive index change in DH GaAs/AlGaAs phase modulators *J. Appl. Phys.* **61**

[26] Mendoza-Alvarez J G, Nunes F D and Pate N B 1980 Refractive index dependence on free carriers for GaAs *J. Appl. Phys.* **51** 4365–4367

[27] Born M and Wolf E 1980 *Principles of Optics* (New York: Pergamon)

[28] Kaminow I P 1974 *An Introduction to Electro Optic Devices* (New York: Academic)

[29] Wakita K, Kotaka I, Asai H, Nojima S and Mikami O 1988 High efficiency electroabsorption in quaternary AlGaInAs quantum well optical modulators *Electron. Lett.* **24** 1324–1326

[30] Ido T, Sano H, Moss D J, Tanaka S and Takai A 1994 Strained InGaAs/InAlAs MQW electroabsorption modulators with large bandwidth and low driving voltage *IEEE Photon. Technol. Lett.* **6** 1207–1209

[31] Devaux F, Dorgeuille F, Ougazzaden A, Huet F, Carenco M, Henry M, Sorel Y, Kerdiles J F and Jeanney E 1993 20 GBit/s operation of a high efficiency InGaAsP/InGaAsP MQW electroabsorption modulator with 1.2 V drive voltage *IEEE Photon. Technol. Lett.* **5** 1288–1290

[32] Wakita K, Kotaka I, Motomi O, Asai H, Kawamura Y and Naganuma M 1990 High speed InGaAlAs/InAlAs multiple quantum well optical modulators *J. Lightwave Technol.* **8** 1027–1032

[33] Wakita K, Yoshino K, Kotaka I, Kondo S and Noguchi Y 1996 Blue chirp electroabsorption modulators with very thick quantum wells *IEEE Photon. Technol. Lett.* **8** 1169–1171

[34] Ido T, Tanaka S, Suzuki M, Koizumi M, Sano H and Inoue H 1996 Ultra high speed multiple quantum well electro absorption optical modulators with integrated waveguides *J. Lightwave Technol.* **14** 2026–2034

[35] Mineo N, Yamada K, Nakamura K, Sakai S and Ushikobo T 1998 60 GHz band electroabsorption modulator module *Optical Fibre Conference* (San Jose, CA, 1998) paper ThH4

[36] Motomi O, Kotaka I, Wakita K, Nojima S, Kawano K, Kawamura Y and Asai H 1992 40 GHz bandwidth InGaAs/InAlAs multiple quantum well optical intensity modulator *Appl. Opt.* **31** 2030–2035

[37] Satzke K, Baums D, Cebulla U, Haisch H, Kaiser D, Lach E, Kuhn E, Weber J, Weinmann R, Widemann P and Zielinski E 1995 Ultrahigh bandwidth (42 GHz) polarization independent ridge waveguide electroabsorption modulator based on tensile strained InGaAsP MQW *Electron. Lett.* **31** 2030–2032

[38] Loi K K, Mei X B, Hodiak J H, Tu C W and Chang W S C 1998 38 GHz bandwidth 1.3 μm MQW electroabsorption modulators for RF photonic links *Electron. Lett.* **34** 1018–1019

[39] Wakita K, Kotaka I, Yoshino K, Kondo S and Noguchi Y 1995 Polarization independent electroabsorption modulators using strain compensated InGaAs/InAlAs MQW structures *IEEE Photon. Technol. Lett.* **7** 1418–1420

[40] Devaux F, Chelles S, Ougazzaden A, Mircea A, Carre M, Huet F, Carenco A, Sorel Y, Kerdiles J F and Henry M 1994 Full polarization insensitivity of a 20 Gb/s strained MQW electroabsorption modulator *IEEE Photon. Technol. Lett.* **6** 1203–1205

[41] Liao H H, Mei X B, Loi K K, Tu C W, Asbeck P M and Chang W S C 1997 Microwave structures for traveling-wave MQW electro-absorption modulators for wide band 1.3 μm photonic links *Optoelectronic Integrated Circuits: Proc. SPIE* vol 3006 (San Jose, CA, USA, 12–14 Feb. 1997) pp 291–300

[42] Zhang S Z, Chiu Y J, Abraham P and Bowers J E 1999 25 GHz polarization insensitive electroabsorption modulators with traveling wave electrodes *IEEE Photon. Technol. Lett.* **11** 191–193

[43] Kawano K, Kohtoku M, Ueki M, Ito T, Kondoh S, Noguchi Y and Hasumi Y 1997 Polarization insensitive traveling wave electrode electroabsorption (TW-EA) modulator with bandwidth over 50 GHz and driving voltage less than 2 V *Electron. Lett.* **33** 1580–1581

[44] Gopalakrishnan G K, Burns W K, McElhanon R W, Bulmer C H and Greenblatt A S 1994 Performance and modeling of broadband $LiNbO_3$ traveling wave optical intensity modulators *J. Lightwave Technol.* **12** 1807–1819

[45] Szentkuti B T 1976 Simple analysis of anisotropic microstrip lines by a transform method *Electron. Lett.* **12** 672–673

[46] Gupta K C, Garg R, Bahl I and Bhartia P 1996 *Microstrip Lines and Slotlines* (Boston: Artech)

[47] Zhang X and Miyoshi T 1995 Optimum design of coplanar waveguide for $LiNbO_3$ optical modulator *IEEE Trans. Microwave Theory Technol.* **43** 523–528

[48] Noguchi K, Mitomi O, Kawano K and Yanagibashi Y 1993 Highly efficient 40-GHz bandwidth Ti:$LiNbO_3$ optical modulator employing ridge structure *IEEE Photon. Technol. Lett.* **5** 52–54

[49] Mitomi O, Noguchi K and Miyazawa H 1995 Design of ultra broad band $LiNbO_3$ optical modulators with ridge structure *IEEE Trans. Microwave Theory Technol.* **43** 2203–2207

[50] Noguchi K, Mitomi O and Miyazawa H 1998 Millimeter-wave Ti:$LiNbO_3$ optical modulators *J. Lightwave Technol.* **16** 615–619

[51] Noguchi K, Miyazawa H and Mitomi O 1998 Frequency dependent propagation characteristics of coplanar waveguide electrode on 100 GHz Ti$LiNbO_3$ optical modulator *Electron. Lett.* **34** 661–663

[52] Gopalakrishnan G K, Burns W K and Bulmer C H 1992 Electrical loss mechanism in traveling wave switch/modulators *Electron. Lett.* **28** 207–209

[53] Burns W K, Howerton M M, Moeller R P, Greenblatt A S and McElhanon R W 1998 Broad band reflection traveling-wave $LiNbO_3$ modulator *IEEE Photon. Technol. Lett.* **10** 805–806

[54] Glass A M 1978 The photorefractive effect *Opt. Eng.* **17** 470–479

[55] Fujiwara T, Sato S and Mori H 1989 Wavelength dependence of photorefractive effect in Ti indiffused $LiNbO_3$ waveguides *Appl. Phys. Lett.* **54** 975–977

[56] Betts G E, O'Donnell F J and Ray K G 1994 Effect of annealing photorefractive damage in titanium indiffused $LiNbO_3$ modulators *IEEE Photon. Technol. Lett.* **6** 211–213

[57] Korotky S K and Veselka J J 1996 An RC network analysis of long term Ti:$LiNbO_3$ bias stability *J. Lightwave Technol.* **14** 2687–2697

[58] Taylor H F 1985 Polarization independent guided wave optical modulators and switches *J. Lightwave Technol.* **3** 1277–1280

[59] Granestrand P, Thylen L and Stoltz B 1988 Polarization independent switch and polarization splitter employing $\Delta\beta$ and $\Delta\kappa$ modulation *Electron. Lett.* **24** 1142–1143

[60] McCaughan L 1984 Low loss polarization independent electrooptic switches *J. Lightwave Technol.* **2** 51–55

[61] Alferness R C 1979 Polarization independent optical directional coupler switching using weighted coupling *Appl. Phys. Lett.* **35** 748–750

[62] Silberberg Y, Perlmutter P and Baran J E 1987 Digital optical switch *Appl. Phys. Lett.* **51** 1230–1232

[63] Bach H G, Kauser J, Nolting H P, Logan R A and Reinhart F K 1983 Electro-optical light modulation in InGaAsP/InP double heterostructure diodes *Appl. Phys. Lett.* **42** 692–694

[64] Sakamoto S R, Jackson A and Dagli N 1999 Substrate removed GaAs/AlGaAs modulator *IEEE Photon. Technol. Lett.* **11** 1244–1246

[65] Walker R G 1991 High speed III–V electrooptic waveguide modulators *IEEE J. Quantum Electron.* **27** 654–667

[66] Walker R G 1995 Electro-optic Modulation at mm-wave frequencies in GaAs/AlGaAs guided wave devices *Proc. IEEE/LEOS'95 8th Annual Meeting* (San Francisco, CA, Oct. 30–Nov. 2, 1995) paper IO4.2, pp 118–119

[67] Spickermann R, Sakamoto S R, Peters M G and Dagli N 1996 GaAs/AlGaAs traveling wave electrooptic modulator with electrical bandwidth greater than 40 GHz *Electron. Lett.* **32** 1095–1096

[68] Spickermann R, Sakamoto S R, Peters M G and Dagli N 1996 GaAs/AlGaAs traveling wave electrooptic modulator with electrical bandwidth greater than 40 GHz *Electron. Lett.* **32** 1095–1096

[69] Sakamoto S, Spickermann R and Dagli N 1995 Novel narrow gap coplanar slow wave electrode for traveling wave electrooptic modulators *Electron. Lett.* **31** 1183–1185

[70] Spickermann R, Peters M and Dagli N 1996 A polarization independent GaAs/AlGaAs electrooptic modulator *IEEE J. Quantum Electron.* **32** 764–769

[71] Spickermann R and Dagli N 1994 Experimental analysis of millimeter wave coplanar waveguide slow wave structures on GaAs *IEEE Trans. Microwave Theory Technol.* **42** 1918–1924

[72] Hornak L A (ed) 1992 *Polymers for Lightwave and Integrated Optics* (New York: Dekker)

[73] Kalluri S, Ziari M, Chen A, Chuyanov V, Steir W H, Chen D, Jalali B, Fetterman H R and Dalton L R 1996 Monolithic integration of waveguide polymer electrooptic modulators on VLSI circuitry *IEEE Photon. Technol. Lett.* **8** 644–646

[74] Kane C F and Krchnavek R R 1995 Benzocyclobutene optical waveguides *IEEE Photon. Technol. Lett.* **7** 535–537

[75] Fishbeck G, Moosburger R, Kostrzewa C, Achen A and Petermann K 1997 Single mode optical waveguides using a high temperature stable polymer with low losses in the 1.55 μm range *Electron. Lett.* **33** 518–519

[76] Lipscomb G F, Garito A F and Narang R S 1981 A large linear electro-optic effect in a polar organic crystal 2-methyl-4-nitroaniline *Appl. Phys. Lett.* **38** 663–665

[77] Chen D, Fetterman H, Chen A, Steier W H, Dalton L R, Wang W and Shi Y 1997 Demonstration of 110 GHz electrooptic polymer modulators *Appl. Phys. Lett.* **70** 3335–3337

[78] Chen D, Bhattacharta D, Udupa A, Tsap B, Fetterman H, Chen A, Lee S S, Chen J, Steier W H and Dalton L R 1999 High frequency polymer modulators with integrated finline transitions and low V_π *IEEE Photon. Technol. Lett.* **11** 54–56

[79] Teng C C 1992 Traveling wave polymeric intensity modulator with more than 40 GHz 3 dB electrical bandwidth *Appl. Phys. Lett.* **60** 1538–1540

[80] Girton G D, Kwiatkowski S L, Lipscomb G F and Lytel R S 1991 20 GHz electrooptic polymer Mach–Zehnder modulator *Appl. Phys. Lett.* **58** 1730–1732

[81] Wang W, Shi Y, Olson D J, Lin W and Bechtel J 1999 Push pull poled polymer Mach Zehnder modulators with a single microstrip line electrode *IEEE Photon. Technol. Lett.* **11** 51–53

[82] Tumolillo T T and Ashley P R 1992 A novel pulse poling technique for EO polymer waveguide devices using device electrode poling *IEEE Photon. Technol. Lett.* **4** 142–145

[83] Hahn K H, Dolfi D W, Moshrefzadeh R S, Pedersen P A and Francis C V 1994 Novel two arm microwave transmission line for high speed electrooptic polymer modulators *Electron. Lett.* **30** 1220–1222

[84] Shi Y, Olson D J and Bechtel J 1996 Photoinduced molecular alignment relaxation in poled electrooptic polymer thin films *Appl. Phys. Lett.* **68** 1040–1042

[85] Mortazavi M, Yoon H and Teng C 1993 Optical power handling properties of polymeric nonlinear optical waveguides *J. Appl. Phys.* **74** 4871–4876

[86] Shi Y, Wang W, Lin W, Olson D J and Bechtel J H 1997 Double end cross linked electrooptic polymer modulators with high optical power handling capability *Appl. Phys. Lett.* **70** 1342–1344

[87] Shi Y, Wang W, Lin W, Olson D J and Bechtel J H 1997 Long term stable direct current bias operation in electrooptic polymer modulators with an electrically compatible multilayer structure *Appl. Phys. Lett.* **71** 2236–2238

[88] Park H, Hwang W Y and Kim J J 1997 Origin of direct current drift in electrooptic polymer modulator *Appl. Phys. Lett.* **70** 2796–2798

B5
Optical amplifiers

Johan Nilsson, Jesper Lægsgaard and Anders Bjarklev

B5.1 Background

From the first demonstrations of optical communication systems, a primary drive in the research activity has been directed towards constant increase in system capacity. Over the past 25 years, it has been an alternating activity to overcome the fundamental fibre limitations of either attenuation or dispersion, and on the basis of the limiting term of the transmission link, systems have been denoted as either *loss limited* or *dispersion limited*. In the mid-1980s the international development had reached a state at which not only dispersion-shifted fibres were available but also spectrally pure signal sources emerged. The long-haul optical communication systems were, therefore, clearly loss limited and their problems had to be overcome by periodic regeneration of the optical signals at repeaters applying conversion to an intermediate electrical signal, which was a complex technology with lack of flexibility.

The technological challenge was to develop a practical way of obtaining the needed gain, that is, to develop relatively simple and flexible optical amplifiers, which would be superior to the electrical regenerators. Several means of doing this had been suggested in the 1960s and 1970s, including direct use of the transmission fibre as amplifying medium through nonlinear effects [1], semiconductor amplifiers with common technical basis in the components used for signal sources [2], or doping optical waveguides with an active material (rare-earth (RE) ions) that could provide the gain [3]. First, however, with the pronounced technological need for optical amplifiers and the spectacular results on erbium-doped fibre amplifiers (EDFAs) [4], an intense worldwide research activity on optical amplifiers was initiated. This resulted in the appearance of commercially available packaged EDFA modules in 1990.

EDFAs operate at signal wavelengths around 1550 nm, i.e. in the so-called third optical transmission window of silica fibres used in telecommunication systems. With the availability of efficient optical amplifiers in this wavelength region, which allowed for cost-effective transmission of optical signals (almost independent of modulation format and bit-rate), the development of multi-channel—wavelength-division-multiplexed (WDM)—systems became possible and a very rapid development of improved amplifier technologies has followed over the past 10 years. EDFAs have truly revolutionized optical telecommunications. By contrast, other optical amplification technologies such as Raman amplification, semiconductor amplifiers, planar amplifiers, and various RE-doped amplifiers for different wavelength regimes have yet to make a significant impact on deployed optical communication systems.

In this chapter, we will provide a short review of the general properties and limitations of optical amplifiers. Thereafter, the specific technologies available for the realization of optical amplifiers are reviewed, and their key parameters are compared. However, since the area of optical amplifiers is extremely broad, we do not attempt to cover all details of the latest development; rather, a selection of recent research and development results will be presented. This selection includes an overall discussion of amplification bands, cladding-pumped amplifiers for high-power amplification, recent results on high-concentration erbium doping in silica glass, and planar optical amplifiers in silica-on-silicon technology.

By and large, we restrict our discussion to amplifiers used in optical transmission systems. There are many other uses for optical amplifiers, in other types of system for other types of application. While all optical amplifiers used in optical transmission systems are waveguiding ones (fibre amplifiers in particular), there are also nonguiding ('bulk') amplifiers. These are primarily used for high-power amplification and amplification of high-energy pulses. Bulk amplifiers have many disadvantages compared to waveguiding ones, including a low gain efficiency and generally a low gain. Furthermore, in the case of a crystalline host, the bandwidth is narrow. Bulk amplifiers are not used in optical transmission systems and will not be discussed here. See, for example, reference [31] for further details.

B5.2 General amplifier concepts

In this section, we discuss some basic concepts common to all optical amplifiers. Due to the general nature of this first part of the chapter, this first part will follow the description given in [13].

Ideally, an optical amplifier would amplify the signal by adding, in phase, a well-defined number of photons to each incident photon. This means that a bit sequence (or analogue optical signal) simply would increase its electromagnetic field strength, but not change its shape by passage through the optical amplifier. Figure B5.1 shows the basics of optical amplification. In a perfect amplifier, this process would take place independent of the wavelength, state of polarization, intensity, (bit) sequence, and optical bandwidth of the incident light signal, and no interaction would take place between signals, if more than one signal were amplified simultaneously. In practice, however, the optical gain depends not only on the wavelength of the incident signal, but also on the electromagnetic field intensity, and thus the power, at any point inside the amplifier. Details of wavelength and intensity dependence of the optical gain depend on the amplifying medium.

An amplifying medium needs to be pumped. For example, a sufficient number of erbium ions must be excited so that a population inversion is reached, otherwise the erbium ions will attenuate rather than amplify the beam. All amplifiers described here need to be optically pumped, with the exception of semiconductor optical amplifiers. Laser diodes are the only acceptable pump sources for telecom applications, because of their efficiency, reliability, and small size. The development of suitable pump diodes has therefore been an integral part of the development of optical amplifiers. However, this will not be discussed here.

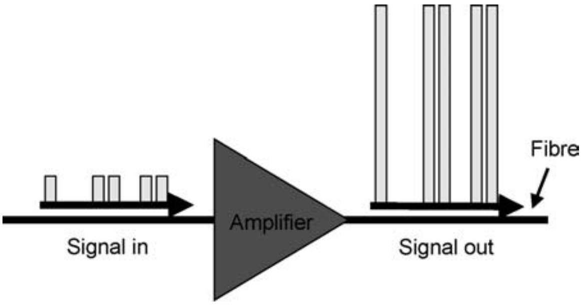


Figure B5.1. Optical amplification. In this case, the amplifier is fibre coupled, so that the input and output signals propagate in an optical fibre.

Next, some general concepts of practical optical amplifiers will be outlined. First, we consider a case in which the amplifying medium is modelled as a homogeneously broadened two-level system. In such a medium, the gain coefficient (i.e. the gain per unit length) can be written as [5]:

$$g(\omega) = \frac{g_0}{1 + (\omega - \omega_0)^2 T_2^2 + P_s/P_{sat}}. \tag{B5.1}$$

Here g_0 is the peak value of the gain coefficient determined by the pumping level of the amplifier, and ω is the optical angular frequency of the incident signal, related to its vacuum wavelength λ by $\omega = 2\pi c/\lambda$. Furthermore, ω_0 is the atomic transition angular frequency. P_s is the optical power of the signal and P_{sat} is the saturation power, which depends on amplifying medium parameters such as fluorescence lifetime and the transition cross section at the signal frequency. The parameter T_2 in equation (B5.1) is normally denoted as the dipole relaxation time [5].

Two important amplifier characteristics are described in equation (B5.1): First, if the signal power ratio obeys $P_s/P_{sat} \ll 1$ throughout the amplifier, the amplifier is said to be operated in the unsaturated region. The gain coefficient is then maximal when the incident angular frequency ω coincides with the atomic transition angular frequency ω_0. The gain reduction for angular frequencies different from ω_0 is generally given by a more complex function than the Lorentzian profile, but this simple example allows us to define the general property of the gain bandwidth. This is normally defined as the full width at half maximum (FWHM) value of the gain coefficient spectrum $g(\omega)$. For the Lorentzian spectrum, the gain bandwidth is given by $\Delta\omega_g = 2/T_2$ (figure B5.2).

From a communication system point of view, it is more natural to use the related concept of amplifier bandwidth (often determined by the 3 dB points) instead of the gain bandwidth, which is related to a point within the amplifying medium. The difference becomes clear, when we consider the linear gain of the amplifier G defined as

$$G = \frac{P_s^{out}}{P_s^{in}} \tag{B5.2}$$

where P_s^{in} is the input power and P_s^{out} is the output power of a continuous wave (cw) signal being amplified. The amplifier gain G may be found by using the relation

$$\frac{dP}{dz} = gP \tag{B5.3}$$

where $P(z)$ is the optical power at a distance z from the amplifier input end. If the gain coefficient $g(\omega)$, for simplicity, is considered constant along the amplifier length, the solution of equation (B5.3)

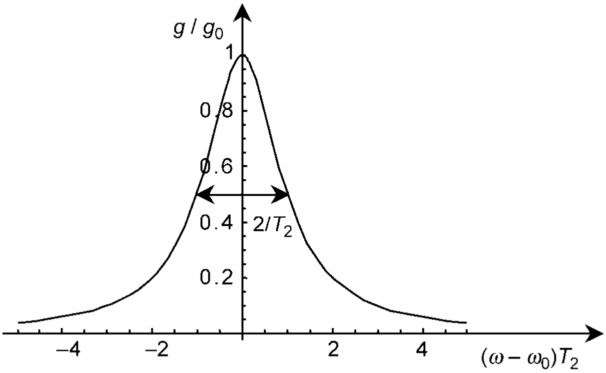


Figure B5.2. Lorentzian gain profile with bandwidth $\Delta\omega_g = 2/T_2$ (FWHM, full width at half maximum). The unsaturated gain is shown, with $P_s \ll P_{sat}$.

is an exponentially growing signal power $P(z) = P_s^{in}\ \exp(gz)$. For an amplifier length L, we then find that

$$G(\omega) = \exp[g(\omega)L]. \tag{B5.4}$$

Equation (B5.4) illustrates the frequency dependence of the amplifier gain G, and shows that $G(\omega)$ decreases much faster than $g(\omega)$ with the signal detuning $(\omega - \omega_0)$, because of the exponential dependence of G on g. Having said that, in many cases the gain spectrum is tophat-like (e.g. supergaussian) rather than Lorentzian, and then, the difference between gain bandwidth and amplifier bandwidth, which is smaller, is reduced. The limited amplifier bandwidth results in signal distortion in the cases where a broadband optical signal is transmitted through the amplifier. In practical optical transmission systems, a very important consequence of a restricted bandwidth relates to WDM signal amplification: different signals at different wavelengths are amplified by different amounts if the amplifier bandwidth is too narrow. The deviation from a flat gain is referred to as gain ripple. The problem of gain ripple is exacerbated in a long-distance transmission system employing many cascaded amplifiers. Any systematic differences in amplification of the different channels then add up, leading to high requirements on gain flatness. To meet this requirement many amplifiers incorporate gain-flattening filters that equalize the gain over a range of wavelengths [38]. While most gain-flattening filters are static, dynamic ones have been developed for demanding applications. In any case, the bandwidth of optical amplifiers is extremely large and one of their principal attractions, maybe around 1 THz even in the simplest amplifier and exceeding 10 THz in state-of-the art, gain-flattened, ones.

Another limitation of the nonideal amplifier is expressed in the power dependence of the gain coefficient. This property, which is known as gain saturation, is included in the example of equation (B5.1), and it appears in all cases where the term P_s/P_{sat} is non-negligible. Since the gain coefficient is reduced when the signal power P_s becomes comparable to the saturation power P_{sat}, the amplification factor (or amplifier gain) G will also decrease. This limits an amplifier's maximum signal output power, known as the saturated output power. Gain saturation might be seen as a serious limitation for multi-wavelength optical communication systems. However, gain-saturated amplifiers have practical use because of their self-regulating effect on the signal output power in links of many concatenated amplifiers. Furthermore, the saturated output power of modern EDFAs is high enough not to be a practical limitation.

Besides the bandwidth and gain saturation limitations of practical optical amplifiers, there is another important limitation in practical amplified systems. Optical amplifiers, in general, will add spontaneously emitted or scattered photons to the signal during the amplification process, and this will consequently lead to a degradation of the signal-to-noise ratio (SNR). The SNR degradation is quantified through a parameter F, normally denoted as the amplifier noise figure, which is defined as the SNR ratio between input and output [6]:

$$F = \frac{\mathrm{SNR_{in}}}{\mathrm{SNR_{out}}}. \tag{B5.5}$$

It should be noted that it is common practice to refer the SNR to the electrical power generated when the optical signal is converted to electrical current by using a photodetector. The noise figure as defined in equation (B5.5), therefore, in general, would depend on several detector parameters, which determine the shot noise and thermal noise associated with the practical detector. However, the influence of detector parameters will not help us to clarify the amplifier noise properties, and it is, consequently, advantageous to consider an ideal detector, whose performance is limited by shot noise only [6].

In practice, the spontaneous emission is reduced by optical filtering of the received signal, by rejecting light at frequencies away from the signal. Therefore, the SNR generally also will be dependent on the bandwidth of the optical filters and the spectral power distribution of the spontaneous emission from the amplifier. However, since this filtering is a process independent of the amplifier properties, it is also common practice to eliminate this ambiguity by considering an ideal filter [5]. Such an ideal filter is introduced only to allow the signal and the spontaneous emission within the signal bandwidth to pass to the detector. Therefore, it will only be the spontaneous emission spectral power density at the signal wavelength that enters the ideal detector, and the noise figure becomes independent of the spectral shape of the spontaneous emission. Note that since the amplifier deteriorates the SNR, that is, $\mathrm{SNR}_{\mathrm{in}} > \mathrm{SNR}_{\mathrm{out}}$, the noise figure will always obey the relation $F > 1$. In fact, conventional, so-called phase-insensitive, amplifiers like the EDFA always have a noise figure of at least 2 (i.e. 3 dB) in the limit of high gain. This is known as the quantum limit. We emphasize that the common interpretation that the output SNR is always at least 3 dB worse than the input SNR is only correct for high-gain amplifiers when the input SNR is shot-noise limited. Though the coherent state emitted from an ideal laser is shot-noise limited, far from all input signals fulfil this. In particular, it is not true for an input signal that has already been amplified.

The noise contribution of an amplifier can also be understood as arising from a fundamentally stochastic nature of the amplification process (stimulated emission). Absorption can also occur in an amplifier, again in a fundamentally stochastic manner that adds to the noise. In order to minimize the noise contribution of an amplifier, one should maximize the stimulated emission to absorption ratio—preferably there should be no absorption at all.

Though the SNR at the output of the amplifier is degraded, the high power of the amplified signal means that additional noise added further down the transmission line will have less impact. Thus, the amplifier improves the SNR of the transmission line as a whole.

A phase-insensitive amplifier operates independently of the optical phase of the signal. In contrast, its counterpart, the phase-sensitive amplifier, provides amplification that does depend on the phase of the signal. A phase-sensitive amplifier does not necessarily degrade the SNR (i.e. it can have a noise figure of unity) [12]. Phase-sensitive amplifiers are used for squeezed-state amplification in research applications, but are not in common use because of practical difficulties.

For a practical communication system, the amplifier spontaneous emission may have another influence besides that described through the noise figure (i.e. besides the effect that it adds fluctuations to the amplified optical signal power, which are converted to current fluctuations during the photo-detection process). This additional influence is that the spontaneous emission, which is emitted from the amplifier input end, may enter the signal source (a semiconductor laser), where it can result in performance disturbances. Therefore, it is often necessary to include isolation between amplifier and light source to avoid additional noise in the system. Therefore, it must be considered that the optical communication system has to be protected against undesired emission from the amplifier. These may also transmit residual pump power onto the transmitter and/or detector. The actual effect of this will be strongly dependent on the spectral properties of the pump light and whether the signal source or detector is sensitive to such radiation.

Another property that has to be evaluated for the optical amplifier is the polarization sensitivity. A high polarization sensitivity means that the amplifier gain G differs for different polarization states of the input signal. Since optical communication systems normally do not include polarization control, and because the polarization state is likely to vary due to external factors such as mechanical pressure, acoustic waves, and temperature variations, the amplifier polarization sensitivity is an undesired property in most cases. Therefore, for amplifiers that are inherently polarization sensitive, it has been a primary goal to reduce or even eliminate the amplifier output power variation due to changes in the signal input polarization state. Fortunately, the dominating EDFA has a negligible intrinsic polarization dependence of the gain.

All of the limiting properties which we have previously discussed are not surprisingly distinctly different for different types of amplifier. This is also the case for the crosstalk limitation that relates to multi-channel applications of optical amplifiers. In contrast to the ideal case, where all signal channels (or wavelengths) are amplified by the same amount, undesired nonlinear phenomena may introduce inter-channel crosstalk (i.e. the modulation of one channel is affected or modified by the presence of another signal channel). We will return to the more specific nature of these nonlinear phenomena in connection with the discussion of the different optical amplifiers.

Other factors such as polarization mode dispersion, multi-path interference, and nonlinearities may also degrade an amplifier.

The final limiting factors that should be mentioned here are closely related to the physical environment in which the amplifier is placed. These may include sensitivity toward vibrations, radioactive radiation, chemicals, pressure, and temperature. However, since at least amplifiers for optical communication systems generally are placed in controlled environments, we will not go further into a detailed discussion of such properties.

The relative importance of the different limiting factors as just discussed depends on the actual amplifier application. One application is the replacement of electronic regenerators; in such cases the amplifiers are placed at a considerable distance from the transmitter and receiver and they are denoted as in-line amplifiers (figure B5.3(*a*)). The optical amplifier may also be used to increase the transmitter power by placing an amplifier just after the transmitter (figure B5.3(*b*)). Such amplifiers are called power amplifiers, or boosters, because their main purpose is to boost the transmitted power. Long-distance systems may also be improved by the inclusion of the so-called preamplifiers, which are placed just before the receiver (figure B5.3(*b*)). Effectively, they improve the sensitivity of the receiver, so that lower-power signals can be detected, thus increasing the distance over which a signal can be transmitted. Furthermore, optical amplifiers can be used in local area networks in which they can compensate for distribution losses (figure B5.3(*c*)). Thereby, the number of nodes in the networks may be increased. Amplifiers can also compensate for the loss of various components, thus enabling the use of lossy optical components such as optical add–drop multiplexers and switches (figure B5.3(*d*)).

It is interesting to note that the most important parameter of an amplifier, its gain, is not a limiting factor, at least not for EDFAs in optical communication systems. They can easily reach a gain of 10 000 (40 dB). However, because of noise limits and nonlinear limits in the transmission system, there is no need for such high gain in a well-designed system. Thus, EDFAs can be designed to reach the gain required in an optically amplified system relatively easily.

B5.3 Different classes of optical amplifiers

The most important factors that limit the performance of practical optical amplifiers have now been identified. The next task is to investigate the technologies available for obtaining real-world amplifiers. Basically, six different ways of obtaining optical amplification are considered: fibre-Raman and fibre-Brillouin amplifiers, semiconductor optical amplifiers (SOAs), RE-doped fibre amplifiers, RE-doped waveguide amplifiers, and optical parametric amplifiers (OPAs). The aim of the following description is to present the basic physical properties of these amplifiers. The properties of the different amplifiers are summarized in table B5.1.

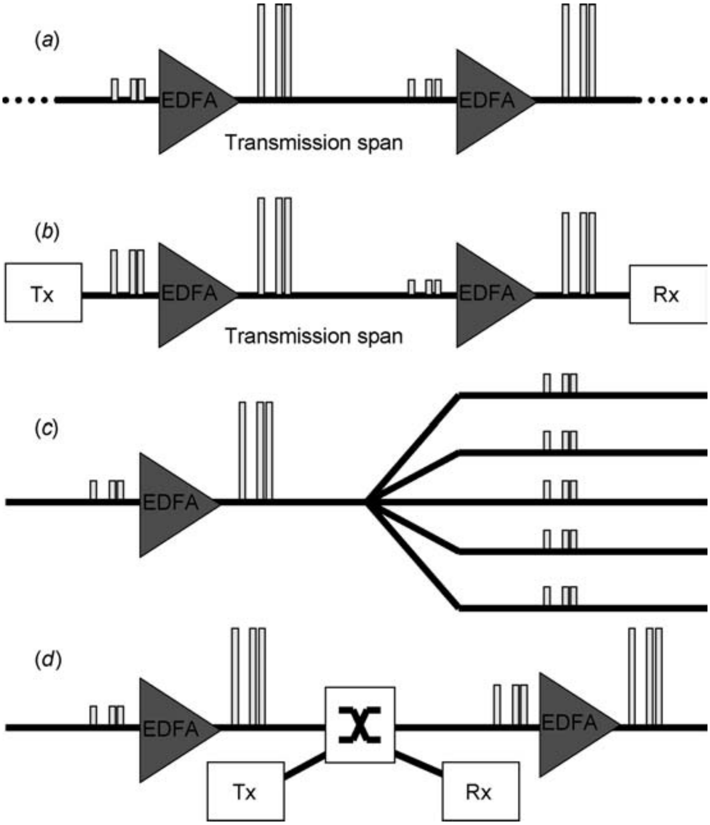


Figure B5.3. Amplification of optical signals with erbium-doped fibre amplifiers (EDFAs) in various optical transmission fibres. (*a*) In-line amplifiers are used periodically in a link comprising several fibre spans. (*b*) A booster amplifier is used after a transmitter (Tx) and a pre-amplifier is used before a receiver (Rx) in a simple repeaterless link. (*c*) A signal is amplified before being split up in a distribution network. (*d*) Amplifiers enable the use of optical add–drop multiplexers and other lossy optical components. Similarly, signals can be amplified in high-power EDFAs to the point where nonlinear optical effects can be utilized, e.g. for all-optical regeneration.

B5.3.1 Fibre-Raman amplifiers

Raman amplification is currently a very active field of research. The fibre-Raman amplifier works through the nonlinear process of stimulated Raman scattering (SRS) occurring in the fibre itself. Like most optical amplifiers, a fibre Raman amplifier must be optically pumped, preferably by a semiconductor laser diode (LD) or a fibre laser. As early as 1962, it was observed that for very intense pump waves a new phenomenon of SRS can occur in which the Stokes wave grows rapidly inside the medium such that most of the pump energy appears in it. SRS is an interaction between light and the vibrational modes of silica molecules that converts a fraction of the incident power from one optical beam (the pump) to another optical beam (the signal, or Stokes wave) at a frequency down-shifted by an amount determined by the vibrational modes (phonon energy) of the medium. A phonon is a quantized vibration of the surrounding medium. SRS involves the so-called optical phonons. Optical phonons of a given energy exist with a broad range of momenta, so that momentum conservation (phase matching) is guaranteed. As a consequence, Stokes waves are generated both co-propagating and counter-propagating with the incident beam. See, e.g. reference [1] for a discussion of SRS.

Table B5.1. Characteristics of optical amplifiers for telecom applications, with an emphasis on commercially available amplifiers.

	EDFA	EDWA	Other RE-doped amplifiers	Fibre Raman	Semiconductor	Fibre OPA	Fibre Brillouin	Comment
Amplifying medium	Er-doped fibre	Er-doped planar waveguide	Rare-earth doped fibre or waveguide	Any fibre; lumped Raman amplifier should use optimized fibre; distributed Raman amplification uses transmission fibre	Semiconductor, e.g. InGaAsP for 1550 nm operation	Dispersion-shifted fibre (zero-dispersion wavelength must coincide with pump wavelength)	Any fibre	
Status, deployment	Widely deployed in commercial systems	Commercially available but not deployed	Some amplifiers commercially available (e.g. Tm and Pr-doped fluoride fibre amplifiers); not deployed	Lumped Raman amplifiers not deployed; small-scale deployment of distributed Raman amplification	Commercially available but not deployed	Not commercially available or deployed; not active R&D area	Not commercially available or deployed; not active R&D area	
Pumping	Optical, at 980 or 1480 nm; typical pump power a few hundred milliwatts	Optical, at 980 or 1480 nm; typical pump power a few hundred milliwatts	Optical; typical pump power a few hundred milliwatts	Optical, e.g. at ~ 1450 nm for amplification at 1550 nm; typical pump power 1 W	Electrical; typical pump current a few hundred milliamperes	Optical, at wavelength near signal wavelength; typical pump power > 1 W	Optical, ~10 GHz above signal frequency	
Gain	20–30 dB typical, 40 dB readily obtainable	20 dB (typical)	20 dB (typical)	15–20 dB optimum for distributed Raman amplification	10–30 dB (typical)	40 dB obtainable; uni-directional if pump is uni-directional	Sufficient; uni-directional if pump is uni-directional	30 dB sufficient for telecom applications; even 10–20 dB is sufficient in many cases

Noise performance	Good (e.g. 4 dB noise figure)	Reasonable (noise figure typically 5–7 dB)	Varying (good in many cases)	Reasonable for lumped RFAs; excellent for distributed Raman amplification	Reasonable (6–10 dB noise figure typical)	Good	Poor	For nonfibre amplifiers, fibre coupling loss degrades the noise figure
Maximum output power	From 10 dBm up to 23–27 dBm in high-power EDFAs, limited by pump power (50 dBm has been demonstrated [63])	Typically 10–15 dBm	Varying	25 dBm (typical)	8–17 dBm (typical)	Typically up to 30 dBm	Depends on pump power	Generally maximum output power depends on pump power, which is high for Raman and parametric amplifiers
Operating wavelengths	1480–1620 nm (C-band 1528–1562 nm, L-band 1570–1620 nm, S-band 1480–1520 nm)	1530–1560 nm	Tm, 1450–1520 nm (S-band); Pr, 1300 nm (2nd telecommunications window); Nd, 1300 nm	Any wavelength, determined by pump wavelength and fibre composition	Any wavelength (determined by bandgap of given material and structure); 1200–1650 nm demonstrated	Any wavelength (near zero-dispersion wavelength of fibre)	Any wavelength, determined by pump wavelength and fibre composition	
Bandwidth	1–10 THz (single band) 8–18 THz (split band)	1–4 THz	1–2 THz	~ 3 THz with single-wavelength pumping, up to ~ 10 THz with broadband pumping; 20 THz possible with tellurite fibres	Up to ~ 10 THz	Typically 1–2 THz (25 THz with pulsed pumping)	Intrinsically below 0.1 GHz	
Polarization dependence of gain	Negligible	Small with appropriate waveguide design	Negligible in fibres (small in appropriately designed planar waveguides)	Small with polarization multiplexed or scrambled pump, otherwise high	Small with appropriate waveguide design (e.g. 0.5 dB)	Intrinsically high, but smaller with a polarization multiplexed or scrambled pump	Small with polarization multiplexed or scrambled pump, otherwise high	

(*continued*)

Table B5.1. (Continued)

	EDFA	EDWA	Other RE-doped amplifiers	Fibre Raman	Semiconductor	Fibre OPA	Fibre Brillouin	Comment
Gain response time	0.1–1 ms	~1 ms	Varying, 0.1 ms typical	Instantaneous	Sub-nanosecond	Instantaneous	Nanosecond	Depends on gain medium, waveguide design, and operating conditions
Extractable energy stored in gain medium	Large	Large	Large	None	Small	None	None	The small energy stored in SOAs and their short response time leads to signal distortion in saturation
Transient behaviour	Transients potentially severe because saturated gain, often suppressed with electronic pump power control	Gain often not saturated, in which case transients are small	Varying	Gain normally not saturated, transients therefore small	OK in special designs (gain-clamped, 'linear', amplifiers)	Gain normally not saturated, transients therefore small		
Gain efficiency	High (a few dB mW^{-1} in C-band)	Reasonable	Varying	Low (up to $\sim 50\,\mathrm{dB\,W}^{-1}$ for distributed Raman amplification in standard single-mode fibre)	High	Low	High	Gain efficiency less important for high-power devices
Power conversion efficiency	Good	Often poor	Varying	Good	Good	Good	Good	Important for high-power devices

WDM capability?	OK	OK	OK	OK	OK in special designs (gain-clamped, ‘linear’, amplifiers)	OK	No	Besides broadband gain, unsaturated operation *or* slow gain dynamics (preferably with transient control) required for WDM amplification
Fibre coupling loss	Low (fusion splicing)	Moderate	Low (splicing)	Low (fusion splicing)	High	Low (fusion splicing)	Low (fusion splicing)	
Typical length of waveguide (or fibre)	1–50 m	0.1–1 m (long waveguides in coiled geometry to reduce device size)	1 m	Kilometre	Millimetre	0.1–1 km (typical)	<1 km	Depends on concentration for RE-doped devices; nonlinear processes occur faster at higher powers, allowing for shorter devices
Cost	Low, and low relative to performance ($1000 EDFAs advertised)	Potentially very low	High	Relatively high	Potentially very low (currently advertised below $1000)	High		Low-cost simpler amplifiers geared towards metro; high-end EDFAs and Raman for long haul
Example of suppliers	Numerous	Inplane Photonics NKT Integration Teem Photonics	NEL	IPG Photonics Licomm MPB Communications Xtera (distributed Raman amplification)	Alcatel Genoa Kamelian Opto Speed	None	None	

In an amplifier context, if the signal angular frequency ω_s and pump angular frequency ω_p are chosen with a difference $\Omega_R = \omega_p - \omega_s$ (the Stokes shift) corresponding to the vibrational energy of the molecules, the signal may experience Raman gain via SRS. Thus, with a proper choice of pump wavelength, Raman amplification is possible at any signal wavelength (e.g. 1300 nm, 1550 nm, ...), limited only by the transparency of the material. Since phase-matching is automatically fulfilled, Raman gain is independent of the pump direction. Another significant feature of the Raman gain in silica fibres is that it extends over a large frequency range (up to 40 THz) with a broad dominant peak near 13 THz with a typical FWHM width of 6 THz (figure B5.4). This is due to the amorphous nature of fused silica in which the molecular vibrational frequencies spread out into bands overlapping each other and creating a continuum (depending on the fibre core composition). The bandwidth of a Raman amplifier would normally be narrower than the intrinsic Raman gain, but it can be extended by using multiple pumps at different wavelengths. Each pump creates Raman gain with a peak down-shifted by ~13 THz. With an appropriate choice of pump wavelengths, the Raman gain spectra of the individual pumps will overlap. Thus, the combined amplification spectrum can be significantly broader, with a low ripple. This is shown in figure B5.5. Thus, Raman amplification bandwidths exceeding 10 THz can be obtained in silica fibres, and with careful optimization, the gain ripple can be below 0.5 dB [54]. In this kind of system there is significant stimulated Raman scattering between the pumps, that redistributes the pump power and therefore needs to be taken into account [57]. However, the pump beams can be time multiplexed in order to suppress the interaction [56]. Also SRS between signals at different wavelengths is a concern in high-power, broadband systems. For bandwidths beyond ~10 THz, the signal and pump beams would have to overlap spectrally with silica fibres. This leads to severe signal degradation, and is why the bandwidth of silica fibre Raman amplifiers is limited to ~10 THz, even though amplification can be obtained at any wavelength with the right pump.

The large bandwidth of fibre Raman amplifiers and a very large saturation power (typically 0.1–1 W) makes them attractive for optical communication systems. Also their noise properties are good, with a noise figure close to 3 dB being obtainable. While stimulated Raman scattering leads to gain, the pump beam will also cause spontaneous Raman scattering. The scattered light is Stokes shifted by the Raman frequency shift, so it will overlap spectrally with the signal. Since spontaneous Raman scattering is a stochastic process, it effectively adds noise to the signal, though the noise addition can be close to the 3 dB quantum limit.

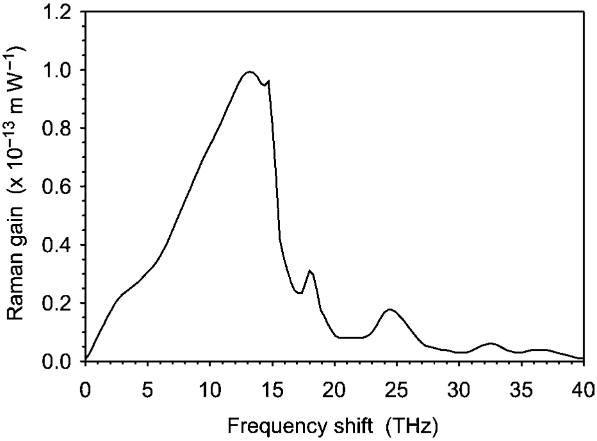


Figure B5.4. Raman gain in fused silica for parallel polarizations, with a pump wavelength of 1064 nm. The Raman gain coefficient scales approximately inversely with the pump wavelength. After [65].

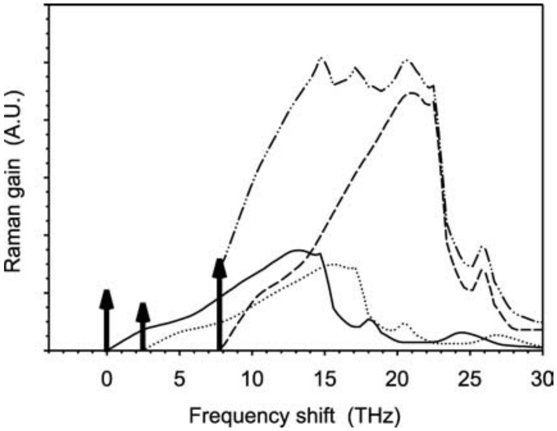


Figure B5.5. Schematic illustration of how several pumps (three in this case) at different wavelengths combine to create a broadened Raman gain.

SRS requires long fibres and high pump powers because it is a weak process. Starting with the Raman gain coefficient shown in figure B5.4, the actual Raman gain can be obtained by dividing the gain coefficient by the effective area (e.g. 50 μm^2) and multiplying it by the effective fibre length (up to ~20 km). Despite the weakness of SRS, with the present development of powerful solid-state pump lasers, sufficient power levels are obtainable from fibre lasers as well as directly from laser diodes. A pump power of 1 W can be enough to generate a gain of up to 50 dB in standard single-mode fibre. For amplification at 1550 nm, a pump wavelength of ~1450 nm is preferable, and suitable laser diodes have been developed at that wavelength. In addition, since SRS will be more efficient if the pump and signal beams are more tightly focused, efficient Raman amplification is possible in fibres shorter than 10 km with optimized designs. In highly nonlinear fibres such as small-core holey fibres, device lengths can be of the order of 100 m.

The weak nature of SRS is actually an advantage for distributed Raman amplification: since Raman gain takes place directly in a silica fibre, it can also occur in the transmission fibre itself. Thus, it is possible to turn the transmission fibre itself into an amplifier, and with an appropriate pump source, it is possible to distribute the gain over several tens of kilometres of transmission fibre. That way amplification can be realized many kilometres away from the pump injection point (often at the receiver end of a transmission link). This leads to a better SNR, since the signal starts to be amplified before it has experienced the full loss of the transmission link (in the absence of Raman gain). Most of the recent record-breaking transmission experiments have indeed utilized distributed Raman amplification in the transmission line [58].

Raman amplification is limited by Rayleigh scattering. At too high gain multiple Rayleigh scattering leads to multi-path interference, which degrades the signal. Optimum distributed Raman on–off gain is in the range 15–20 dB.

Polarization is another concern for Raman amplification. The gain in a polarization orthogonal to the pump polarization is only a few percent of the gain for parallel polarization [53,66,67]. Thus, the polarization properties must also be taken into account when evaluating the Raman gain in a fibre [67]. Because polarization control is not normally employed in optical communication systems, this property introduces an additional complexity. In practice, the polarization dependence can be mitigated with pump polarization scrambling or multiplexing.

The numbers quoted so far refer to silica and germanosilicate fibres. Recently, however, Raman amplifiers have been demonstrated in tellurite fibres. Tellurite is a nonsilica glass with a significantly broader bandwidth. An amplification bandwidth of 20 THz, from 1490 to 1650 nm, has so far been demonstrated, which is a record for a single-band cw amplifier [52].

B5.3.2 Brillouin amplifiers

Because of severe limitations such as narrow linewidth, Brillouin fibre amplifiers are not practical and will not be treated henceforth. See, e.g. reference [1] for details on Brillouin amplifiers.

B5.3.3 Semiconductor optical amplifiers

The SOA is fabricated in a material and structure very different from the silica transmission fibre. It incorporates a waveguide within a semiconductor gain medium. For 1550 nm amplification, InGaAsP SOAs are often used. The major attraction of SOAs is the potential for low cost and the direct electrical pumping.

Since an SOA is not a fibre device, the coupling of light from fibre to SOA and vice versa is difficult. There are inevitably significant coupling losses which degrade the noise figure. Furthermore, intrinsically, the SOA itself also will experience relatively large feedback due to reflections occurring at the cleaved end facets (32% reflectivity due to refractive index differences between semiconductor and air), resulting in sharp and highly disadvantageous gain reduction between the cavity resonances of the Fabry–Perot resonator. Therefore, it is generally necessary to design travelling wave (TW) type SOAs by suppressing the reflections from the end facets; a common solution is the inclusion of antireflection coating of the end facets. It turns out that to avoid the amplifier bandwidth being determined by the cavity resonances rather than the gain spectrum itself, it becomes necessary to require that the facet reflectivities satisfy the condition [5]:

$$G\sqrt{R_1R_2} < 0.17. \tag{B5.6}$$

Here, G is the single-pass amplification factor, and R_1 and R_2 are the power reflection coefficients at the input and output facets, respectively. For an SOA designed to provide a 30 dB gain, this condition will mean that $\sqrt{R_1R_2} < 0.17 \times 10^{-3}$, which is difficult to obtain in a predictable and reproducible manner by antireflection coating alone. Additional methods to reduce the reflection feedback include designs in which the active-region stripe is tilted from the facet normal and introduction of a transparent window region (nonguiding) between the active-layer ends and the antireflection coated facets. Thereby, reflectivities as small as 10^{-4} may be provided, so that the SOA bandwidth can be determined by the amplifying medium rather than by the narrower resonance peaks of the cavity. Typically, 3 dB amplifier bandwidths of 60–80 nm may be obtained. Another attractive spectroscopic property of SOAs is that, with appropriate composition and design, gain can be realized over a very wide range of wavelengths from 1200 to 1650 nm (though limited to, say, 80 nm in a given device with a specific bandgap).

The gain in an SOA depends on the carrier population, which changes with the signal power and the injection current. To reiterate, the SOA is pumped electrically. By contrast, alternative optical amplifiers are optically pumped. Typical pump currents are a few hundred milliamperes. An important property is the very short effective lifetime of the injected carriers (e.g. of the order of 100 ps). This property becomes specifically relevant for multi-channel applications of SOAs, where crosstalk limitations arise. This crosstalk originates from two nonlinear phenomena: cross-saturation and four-wave mixing. The former appears because the signal in one channel through stimulated recombinations affects the carrier population and, thereby, the gain of other channels. This significant problem may only be reduced by operating the amplifier well below saturation, but this is not an easy task due to the relatively limited saturation output power of the order of 10 mW. Four-wave mixing also appears because stimulated recombinations affect the carrier number. More specifically, the carrier population may be found to oscillate at the beat frequencies between the different channels, whereby both gain and refractive index are modulated. The multi-channel signal, therefore, creates gain and index gratings, which will introduce inter-channel crosstalk by scattering a part of the signal from one channel to another.

Despite these difficulties, the so-called gain-clamped (‘linear’) SOAs for multi-wavelength amplification have been realized [36, 37]. Because of their linear (power-independent) characteristics, channel cross talk and four-wave mixing can be avoided. For instance, a linear SOA with 10 dBm output power, up to 25 dB gain, and 8 dB noise figure is offered commercially by Genoa [32].

As with other types of optical amplifier, the stimulated emission is accompanied by spontaneous emission, which leads to noise. The noise properties of the SOA are determined by two factors. One is the emission due to spontaneous decays and the other is the result of nonresonant internal losses α_{int} (e.g. free carrier absorption or scattering loss), which reduce the available gain from g to $(g - \alpha_{\mathrm{int}})$. The nonzero absorption then increases the noise figure. Also residual facet reflectivities increase the noise figure, via loss of signal input power and via multi-path interference. Typical values of the noise figure for SOAs are 6–10 dB.

Another undesirable characteristic of early SOAs is the polarization sensitivity, which appears because the amplifier gain differs for the transverse-electric (TE) and transverse-magnetic (TM) modes in the semiconductor waveguide structure. However, intense research in the early 1990s has reduced the problem, and SOAs with polarization sensitivity reduced to less than 0.5 dB have been reported [7]. Other methods of using serial or parallel coupled amplifiers or two passes through the same amplifier are also suggested. Although such schemes increase complexity, cost, and stability requirements, they may add attractive properties applicable in optical signal processing (e.g. in optical wavelength conversion [7]).

To conclude the discussion of the SOA, it is important to note that the major drawbacks are polarization sensitivity, interchannel crosstalk, and large coupling losses. These drawbacks have been overcome to some extent, but the resulting performance is still significantly worse than that of EDFAs. Therefore, the impact of SOAs on real systems has been small so far. In favour of the SOA are the large amplifier bandwidth, the possibility of amplification at any wavelength in the range 1200–1650 nm (at present), and the possibility of monolithic optoelectronic integration, especially within the receiver, where the input signal powers are weak enough to avoid undesired nonlinearities. It should also be mentioned that the SOA itself can be used as an amplifier and a detector at the same time, because the voltage across the pn-junction depends on the carrier density, which again interacts with the optical input signal. Promising results for transparent channel drops, channel monitoring, and automatic gain control have been demonstrated [7]. Note that possible future applications of SOAs are not limited to amplification alone; they can also be used as wavelength conversion elements and optical gates [7].

B5.3.4 Rare-earth-doped fibre amplifiers

Parallel to the maturation of SOAs, another development has taken place, with a revolutionary impact on optical communication systems. With a point of reference in work on RE-doped glass lasers initiated as early as 1963 [3], the first low-loss, single-mode, RE-doped fibre amplifiers (as possible useful devices for telecommunication applications) were demonstrated in 1987 [4]. Progress since then has multiplied to the extent that amplifiers today offer far-reaching new opportunities in telecommunication networks. We will in the following sections discuss some of the recent results on RE-doped fibre amplifiers.

The possible operational wavelengths of RE-doped fibre amplifiers are determined by the emission spectra of the RE ions, moderately dependent on the host material in which they are embedded. Only a few RE materials become relevant for optical communication purposes, primarily erbium, which may provide amplification in the 1550 nm wavelength band (the third telecom window), and praseodymium, which may be operated around the 1300 nm band (the second telecom window). Note that also ytterbium and thulium have shown important features in amplifier development over the past few years.

The absorption spectrum holds accurate information about the location of possible pump wavelengths that can be used to excite the RE ions to higher energy levels. From this higher energy level the ion can relax to the ground state, transferring its packet of energy either radiatively or nonradiatively. The most prominent nonradiative de-excitation mechanism, known as multi-phonon relaxation, involves the creation of phonons. The maximum energy of a phonon is limited in any given host, for example, in silica to approximately 1100 cm^{-1} (corresponding to a frequency of 33 THz). Thus, the larger the energy gap that is to be bridged via multi-phonon relaxation, the more phonons need to be created. However, this becomes an increasingly improbable process. At energy gaps exceeding around six times the maximum phonon energy, multi-phonon relaxation becomes insignificant. Then, other processes, notably radiative decay, will dominate the relaxation process. Figure B5.6 shows the energy levels, possible pumping wavelengths (980 and 1480 nm), and emission processes for erbium. Quite fortuitously, the phonon energy in silica is such that the 980 nm pump level ($^4\mathrm{I}_{11/2}$) relaxes via rapid multi-phonon relaxation to the metastable level ($^4\mathrm{I}_{13/2}$), while multi-phonon relaxation from $^4\mathrm{I}_{13/2}$ is virtually absent. Instead, the decay from the meta-stable level to the lower energy level (the ground state $^4\mathrm{I}_{15/2}$) is radiative. Thus, it leads to the emission of a photon, either via spontaneous or stimulated emission. Spontaneous emission always takes place, when an amplifying medium such as a collection of ions is in an excited state; therefore, spontaneously emitted light cannot be avoided in a fibre amplifier. Stimulated emission is the process that allows signal amplification to take place and, therefore, is the desired property of the fibre amplifier. The process may be explained as follows: a photon incident on the medium, with an energy equal to the difference in energy of the ground state and an excited metastable state, promotes de-excitation, with the creation of a photon that is in phase with the incident photon.

Radiative transitions within triply ionized RE ions are actually relatively weak because they are ‘forbidden’ for reasons of symmetry (the normally dominant electrical dipole transition is not allowed for the relevant intra-4f transitions). Because of the weak nature, the involved lifetimes in the upper laser level (a measure of the spontaneous decay) are many orders of magnitude larger than the nonradiative lifetimes of the higher energy levels, and many orders of magnitude larger than the radiative lifetime of states that can decay via electric dipole transitions (typical, e.g. for transition metals). In Er^{3+}, the upper-level lifetime is typically 10–14 ms.

The absorption and emission cross-sections for Er^{3+} are shown in figure B5.7. Their size is the measure of the probability for a photon to interact with an Er^{3+} ion in the ground state or metastable

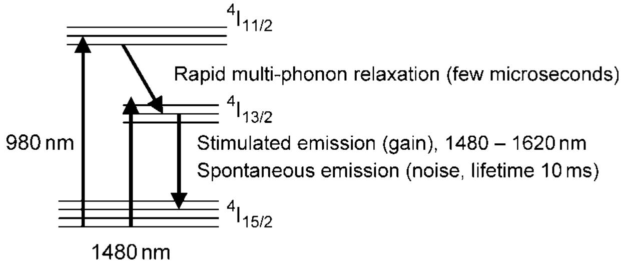

$^4\mathrm{I}_{11/2}$
Rapid multi-phonon relaxation (few microseconds)
$^4\mathrm{I}_{13/2}$
980 nm
Stimulated emission (gain), 1480 – 1620 nm
Spontaneous emission (noise, lifetime 10 ms)
$^4\mathrm{I}_{15/2}$
1480 nm

Figure B5.6. Energy levels of Er^{3+}. Each level is further split by the so-called Stark effect, as schematically illustrated. Within each level, the Stark sub-levels are in thermal equilibrium.

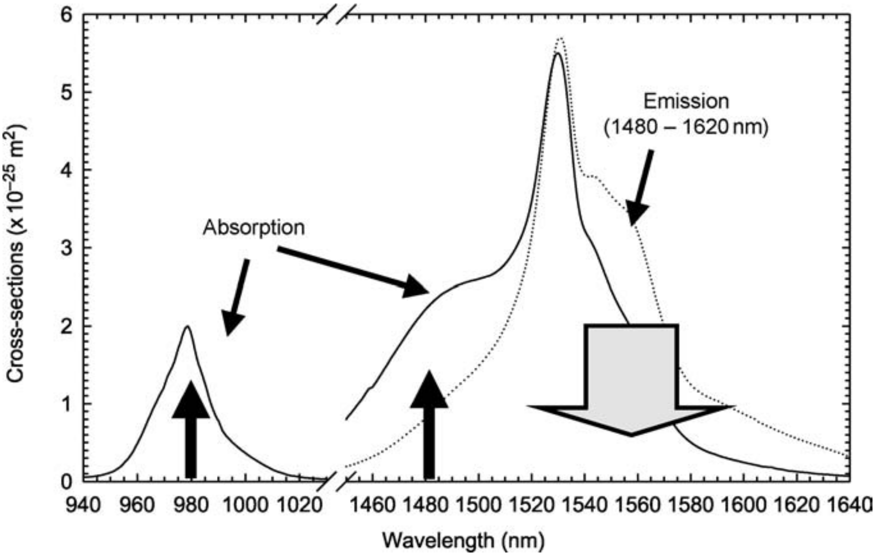


Figure B5.7. Absorption cross-sections (from the ground state) and emission cross-sections (from the metastable state) for Er^{3+}. Possible pump wavelengths and emission bands are indicated.

state. They are wavelength dependent, and allow absorption and stimulated emission in Er^{3+} to be described quantitatively.

EDFAs are commercially available and are widely deployed. For this reason, we will describe the EDFA in some detail.

EDFAs have revolutionized optical communication by removing the loss limit from optical transmission systems, enabling simultaneous amplification of a large number of optical channels, and facilitating the use of lossy components, as well as nonlinear components that only work at high powers. Hence, more complex links and networks with vastly higher capacity are made possible. As shown in figure B5.7, EDFAs operate in the third transmission window around 1550 nm. Figure B5.8 illustrates a simple EDFA. An input signal is launched into the erbium-doped fibre via an isolator and a wavelength-selective coupler. A pump beam from a pump diode is combined with the signal beam in the coupler, and also launched into the erbium-doped fibre. There, it is absorbed by the erbium ions, which thereby are excited to the metastable state. Thus, the erbium-doped fibre can amplify the signal via stimulated emission. The amplified signal exits the amplifier through a second isolator. The isolators only transmit

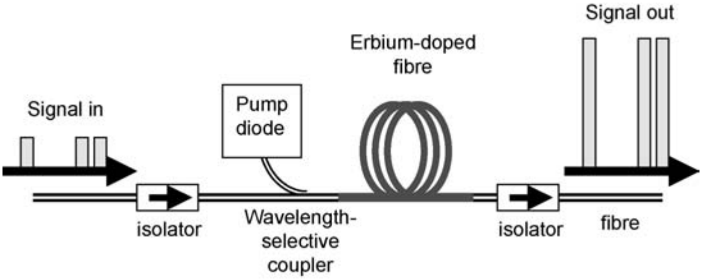


Figure B5.8. Schematic of an erbium-doped fibre amplifier.

light in one direction, and protect the amplifier from external feedback from reflections. Though the isolators and pump diode are discrete components, they are all fibre pigtailed so that the amplifier can be fusion-spliced together.

Real EDFAs are more complicated than the simple one shown in figure B5.8. They often consist of two amplifier stages, and may incorporate a gain-equalizing filter (figure B5.9). Furthermore, the input and output signal powers are monitored, and the pump power is controlled to maintain a constant gain. The pump power control must be fast enough to suppress transients that can arise in amplified systems and destroy them. Telemetry systems need to be implemented, too. For demanding applications, the erbium-doped fibre may be temperature controlled.

An EDFA may be fabricated using a silica glass host, and semiconductor pump sources have been successfully used to pump amplifiers in the 800, 980, and 1480 nm absorption bands. In addition, pumping at shorter wavelengths is also possible, but less interesting due to the lack of practical pump sources [8]. Basic differences exist between the application of the three mentioned pump choices. First, it should be noted that amplification occurs according to a three-level scheme when 800 or 980 nm pumping is applied, but the erbium ion works as a two-level system when 1480 nm pumping is used. Furthermore, 800 nm pumping is much less efficient than 980 nm pumping due to a pronounced excited-state absorption (ESA) of pump photons [8]. Because of these differences, today only 980 and 1480 nm pumping are considered in practical system applications. In cases, where very low noise figures are required, 980 nm pumping is the preferred choice, since the three-level nature of the system makes excitation of essentially all erbium ions (full inversion) possible, resulting in a noise figure very close to 3 dB. For 1480 nm pumping only around 70% of the erbium ions may be excited to the upper laser level. Ions in the ground state absorb the signal and thus degrade the noise figure. Noise figures down to 5 dB are generally obtainable with 1480 nm pumping. However, 1480 nm pumping still provides highly efficient amplifiers.

EDFAs readily provide small-signal gain in the order of 30–40 dB, and values as high as 54 dB have been achieved [9]. The EDFA also provides high saturation powers, and signal output powers of more than 500 mW have been demonstrated; multi-watt output power is even possible in cladding-pumped, ytterbium-sensitized, EDFAs [63]. In such power amplifier applications, 80% or more of the pump photons may be converted to signal photons, indicating the high efficiency of the EDFA. A 30 nm bandwidth of the EDFA may easily be obtained, and is available in commercial amplifiers.

The long lifetime in the upper laser level makes the EDFA an excellent energy reservoir, and it is important to note that, with the power levels used in modern optical communication systems, one signal

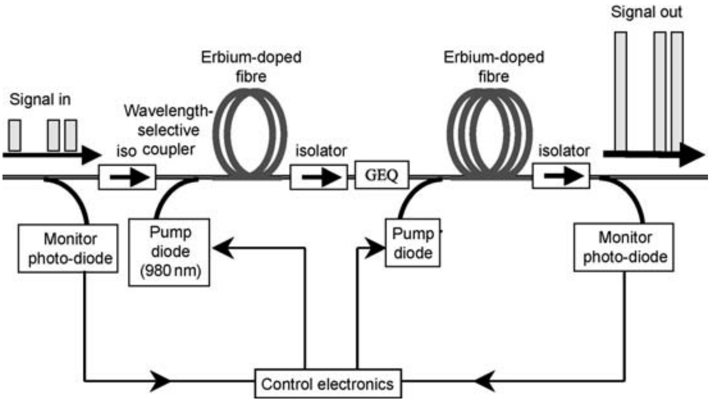


Figure B5.9. Two-stage erbium-doped fibre amplifier with gain-flattening filter (GEQ) and control electronics.

pulse will only interact with a very small fraction of the erbium ions within the EDFA. The following pulse is, therefore, very unlikely to interact with the same erbium ions, and it will remain unaffected by other pulses. For this reason, in practice no crosstalk will be seen in fibre amplifiers. The EDFA is also essentially insensitive to polarization variations of the signal, and only systems with cascaded amplifiers have lately demonstrated degradation due to polarization effects. A typical gain variation of 0.1 dB was found [10], when the polarization state of the signal is switched between orthogonal linearly polarized states. Finally, it should be mentioned that since the EDFA is an optical fibre itself, very low coupling losses (comparable to ordinary splice losses) are involved in the application of EDFAs.

It should be stressed that EDFAs work very well, and well-engineered EDFAs are available commercially. Performance-wise, EDFAs can provide more power, and more gain, than what is required in a telecom system. The noise figure can be very near the quantum limit of 3 dB of traditional (phase-insensitive) amplifiers. There is essentially no polarization dependence. They are very efficient. In these aspects, there is no need, or it is fundamentally impossible, to improve the EDFA.

EDFAs with larger bandwidth would arguably be desirable. Most EDFAs, and all of those used in real systems, are made with silica fibres, often Er-doped aluminosilicate. Therefore, this article is focused on silica EDFAs. Other types of EDFA, made, e.g. with fluoride, tellurite, or bismuth glass may offer some performance advantages, especially in terms of bandwidth. Unfortunately fibres from such glasses are vastly inferior to silica fibres in properties such as durability and strength, and are much more difficult to fabricate and splice. Furthermore fluoride glasses are hygroscopic. In fact, silica EDFAs have been engineered to the point where they outperform other types of EDFA in all important respects, and the over 30 nm bandwidth available in typical broadband commercial silica EDFAs is sufficient for the vast majority of today's transmission systems. Furthermore bandwidths of up to 48 nm have been demonstrated in an antimony silicate EDFA operating at around 1550 nm [50]. This bandwidth is still not quite as large as that of the SOA, but the bandwidth can be extended if one considers other amplification bands: during the first years of the EDFA research and commercial employment, the primary interest was focused on the spectral band from ~1530 to ~1560 nm, because limited lengths of erbium-doped fibres allow for very efficient amplification in this frequency range, and because this range coincides with the loss minimum of silica fibres. Hence, this band has been denoted as the 'conventional' band or just C-band. This use of names naturally indicates that other bands have emerged. Indeed, differently designed EDFAs allow for amplification in the 1570–1620 nm wavelength range, where the loss of silica fibres is still quite low. Such EDFAs are generally said to operate in the long-wavelength or L-band. For example, Codemard *et al* [59] report an unflattened amplification band from 1555 to 1620 nm in a cladding-pumped ytterbium-sensitized phosphosilicate EDFA. L-band amplification can be realized by operating the EDFA at a low inversion level (with only a small fraction of the Er ions excited). This leads to a ground-state absorption in the C-band, which suppresses gain there. At longer wavelengths, the transition becomes quasi-four level (with insignificant ground-state absorption), so a low but finite gain can be realized even with a low inversion level. Long fibres must then be used in order to reach a sufficient gain.

Amplifiers operating on the short-wavelength side of the C-band (from around 1450 to 1520 nm) are defined by the International Telecommunications Union (ITU) to belong to the class of S-band amplifiers. Sometimes the range 1450–1480 nm is called the S^+-band. Transmission in the S-band, at wavelengths between the second and third telecom windows, is possible due to the low loss at those wavelengths in newly developed low-OH silica fibres. High-gain EDFA have also been demonstrated in the S^+-band, down to 1480 nm [51]. In order to realize an S^+-band EDFA one must operate at a high inversion level in order to bleach the large erbium absorption in the S^+-band. Furthermore, a distributed filter is required to suppress the C-band gain, which otherwise saturates the amplifier. Note that S^+-band EDFAs are generally considerably more complicated than C- and L-band EDFAs. All these types of amplifier have been realized with silica fibres.

Thulium-doped fibre amplifiers (TDFAs) are in fact dominating in the S-band while S-band EDFAs are a later development. Typically, TDFAs provide amplification in the 1450–1480 nm window [14], though TDFAs gain-shifted to longer wavelengths within the S-band have been demonstrated. Thus, TDFAs can access the whole S-band. Another option for extending the gain to longer wavelengths is to cascade a TDFA with a discrete Raman amplifier, e.g. to a total bandwidth of 40 nm [15]. Still, compared to (C-band) EDFAs, TDFAs have several disadvantages. The energy-level diagram of triply ionized thulium ions (Tm^{3+}) is quite complicated, and many levels play important roles in S-band TDFAs. One problem is that the upper laser level is quenched in silica fibres by multi-phonon relaxation. Therefore, nonsilica glasses with lower maximum phonon energy such as fluoride and, recently, tellurite glasses are used instead. Another problem is that the lower laser level needs to be depopulated in order to maintain population inversion and avoid the so-called self-termination. Because of this, complicated pumping schemes, including two-wavelength pumping schemes, are employed. Nevertheless, TDFAs are commercially available, and high-gain (20–30 dB), high-output-power (e.g. 20 dBm), and low-noise operation (~5 dB noise figure) has been demonstrated.

Though transmission at around 1550 nm is currently the dominating technology, historically, large-scale deployment of systems in the second transmission window at 1300 nm predates this development. As a result, there has been a large demand for amplifiers at 1300 nm. Unfortunately, RE-doped fibre amplifiers for the second transmission window at 1300 nm have to deal with a number of difficulties, and their applicability is quite limited. The neodymium-doped fibre amplifier is limited by excited-state absorption to wavelengths that are too long (around 1340 nm and beyond) for real systems (at ~1310 nm), and so the best RE candidate in the 1300 nm wavelength band is the praseodymium-doped fibre amplifier (PDFA). The primary difficulty is that also praseodymium has to be placed in a host material other than silica, with a lower phonon energy, since otherwise, the upper energy level of the transition in question will be quenched by multi-phonon relaxation, and no net gain will appear [8]. Until now the most promising host materials have been different compositions of fluoride glasses, with problems such as low strength and hygroscopy. Another problem is that the 1300 nm amplifiers normally have lower pump efficiency than EDFAs. However, pump efficiency as high as 4 dB mW^{-1} for praseodymium-doped amplifiers is shown together with noise figures in the 5–8 dB range.

B5.3.5 Rare-earth-doped waveguide amplifiers

RE-doped waveguide amplifiers are only fundamentally different from fibre amplifiers in the respect that the RE material is embedded in a planar optical waveguide and not a fibre. This means at first glance that only another waveguide geometry has to be considered. However, several important factors have to be considered in the design of RE-doped waveguide amplifiers. First, the background loss is several orders of magnitude larger in a planar waveguide (around 0.01–0.05 dB cm^{-1}) than in a fibre, and second, a much higher RE-dopant concentration is necessary to achieve high gain in the relatively short planar waveguides. This raises the problem of energy transfer between the RE ions and results in a lower amplifier efficiency. Not surprisingly, the focus has been on the erbium-doped planar waveguide amplifiers in silica, and integration with a 980/1530 nm wavelength division multiplexer was presented early in the development [11]. In this component a maximum gain of 13.5 dB for a 600 mW pump power was shown, but gain values up to 20 dB have been demonstrated in other amplifiers. Recent results by Hübner *et al* [64] also demonstrate that boundaries between active and passive regions may be fabricated with transition losses below 0.03 dB/transition and back reflections suppressed more than 70 dB. However, it is not realistic to expect that the RE-doped waveguide amplifier will completely replace EDFAs as high-performance amplifiers in optical communication systems, but their advantages have to be found in the ability for integration of several functionalities, e.g. splitting and amplification on the same component. In addition, planar waveguide amplifiers can be realized in a wider range of materials than fibres. We may consider host materials such as lithium niobate, which is electro-optic, and new applications besides optical communication. Furthermore, Nd-doped planar waveguides in fluoro-aluminate glass remain a possibility for 1310 nm amplification [55]. We will return to more detailed descriptions of properties of Er-doped planar waveguide amplifiers and high-concentration doping in the following sections.

B5.3.6 Optical parametric amplifiers

An OPA utilizes a so-called wavemixing process, in which pump photons at one wavelength are converted directly to signal photons at another wavelength through a lossless nonlinear interaction in a medium. At the same time, idler photons are created in an idler beam at yet another wavelength to satisfy requirements on energy and momentum conservation. It is referred to as a parametric process as it originates from light-induced modulation of a medium parameter such as the refractive index. Its origin lies in the nonlinear response of bound electrons of the material to an applied optical field. Optical parametric amplification is traditionally realized via three-wave mixing (or three-photon mixing) in electro-optic materials such as lithium niobate with a second-order nonlinearity. However, it is also possible in media without a second-order nonlinearity, such as silica [33, 44]. In this case, a third-order nonlinearity (the nonlinear refractive index) leads to amplification. This is now a four-photon process (i.e. a four-wave-mixing process), involving two pump photons, a signal photon, and an idler photon. Two pump photons, from a single or from two different pump beams, are annihilated, and a signal photon and an idler photon are created. Thus, the idler becomes an image of the signal. This allows OPAs to be used for signal wavelength conversion. Because of energy conservation, the signal and idler photons are created at optical frequencies located symmetrically on opposite sides of the frequency (or frequencies) of the pump beam(s). Though four-wave mixing in silica is intrinsically a much weaker process than three-wave mixing in electro-optic materials, the low loss (allowing for long interaction lengths) and tight modal confinement of silica fibres enhances nonlinear interactions and thus compensates for the intrinsic weakness of the four-wave mixing process.

Interestingly, OPAs can be configured both as conventional phase-insensitive amplifiers as well as phase-sensitive ones [34]. In the case of a phase-sensitive fibre OPA, powerful pump as well as idler beams need to be injected into the amplifier. The pump and idler beams interact to create gain for a signal beam of a particular optical phase. Any signal beam with opposite phase would be attenuated, in a process in which a signal and an idler photon combine to create two pump photons. The difficulties in tracking and controlling optical phases are major practical drawbacks of phase-sensitive OPAs.

There is no idler beam injected into a phase-insensitive OPA. Instead, the idler is created in the four-wave mixing process, with an optical phase that depends on the phase of the signal. Because of this, the signal can be amplified independently of its phase, and there is no need to control the optical phase of any beam. Nevertheless, there are many challenges with a phase-insensitive OPA (which also apply to phase-sensitive ones). The main difficulty is that phase matching of the interacting pump, signal, and idler waves has to be maintained along the fibre. This means that fibres with tailored dispersion properties must be used, and the pump wavelength must be near the zero-dispersion wavelength of the fibre. Four-wave mixing depends on the polarization of the interacting waves, so in a typical system without control of signal polarization, the pump beam should be unpolarized, polarization scrambled, or polarization multiplexed. Furthermore, a high-pump-power requirement is a serious limitation. Still, for example, the pump power needed to reach a certain gain with OPAs may be less than half of what is needed in a Raman amplifier. Thus, an OPA has a higher gain efficiency, though its power conversion efficiency may be lower. In addition, an OPA does not require long fibres as a Raman amplifier does. Phase-insensitive fibre OPAs were first demonstrated with 49 dB peak fibre gain for 32.2 dBm of launched pump power [33]. The amplification bandwidth was approximately 10 nm on both sides of the pump wavelength of 1562 nm (parametric gain occurs symmetrically, in wavelength ranges on both sides of the pump wavelength in a phase-insensitive parametric fibre amplifier). The noise figure was inferred to be similar to that of an EDFA. The bandwidth of OPAs increases with pump power, and Ho *et al* recently demonstrated a pulsed fibre OPA with more than 200 nm of total bandwidth using only 20 m of highly nonlinear fibre [35]. High gain, high output power, and low noise figure (e.g. 4 dB) have been confirmed in several recent experiments.

Parametric amplifiers have several attractions, but they are at present not developed to the level of erbium-doped, semiconductor, or Raman optical amplifiers. See, e.g. reference [60] for a recent summary on fibre OPAs.

B5.4 Challenges in fibre amplifier application and development

To reiterate, the performance of EDFAs is sufficient for transmission applications, and there is not much reason to improve performance. A major challenge is instead to cut cost, e.g. by reducing the cost of pump diodes and other EDFA components. A lower cost would increase the proliferation of EDFAs in the network, and include uses for which the specifications are not so demanding. This allows for further cost reductions. Other cost-reduction strategies include a higher degree of parallelization and integration (e.g. by employing planar technology), as well as cladding pumping, both to be discussed further below.

The only performance limitation of EDFAs is its limited wavelength range. Still, the EDFA bandwidth is sufficient for today’s systems. Nevertheless, the limited bandwidth may well become a bottleneck some day. Since the available amplification bandwidth—as indicated—may not be extended significantly beyond the C- or L-band using established EDFA technology, and since the RE-doped S-band amplifiers still are complex (including S-band EDFAs), researchers have begun seeking alternative enabling techniques wider than those that can be managed by erbium- (or even thulium-) doped amplifiers. In this context, broadband, wavelength-agile technologies such as SOAs and Raman amplifiers, discussed above, are attractive options. Also from a cost perspective, SOAs are potentially very attractive, but they still require further development before they can be widely deployed.

Distributed Raman amplification directly in the transmission line can provide the best overall system performance, and offer the longest reach of optical networks. Still, the cost, dominated by the pumping system, must be reduced before Raman-amplified systems will be widely deployed.

B5.5 Cladding-pumping

Early fibre amplifiers were pumped by large laser systems requiring many kilowatts of electrical power, with an overall EDFA power conversion efficiency (‘wallplug efficiency’) of the order of one in a million. EDFAs only became practical with the realization of suitable semiconductor laser diodes for pumping. These pump-LDs improved the wallplug efficiency to the per cent level and above. In most EDFAs, the optical pump beam is injected into the core, which implies that a (transversally) single-mode pump source must be used. Though single-mode pump diodes are being developed rapidly, they are still limited in power to less than a watt. While an EDFA is typically pumped by more than one LD, this power limitation still means that, in practice, traditional core-pumped EDFAs are limited in power to less than 1 W (i.e. 30 dBm). This power is more than enough for most applications in optical communications, but higher powers can be used in free-space communication and in distribution networks in which the transmitted power (e.g. 1 W from an EDFA) is divided and distributed in a large number of different fibre branches.

Cladding-pumping is a method of overcoming the power limitation of core-pumped EDFAs. Cladding-pumped fibre devices utilize multi-mode pump-LDs rather than single-mode ones.

Multi-mode diodes are more powerful and often cheaper than single-mode ones. They come in many shapes for a wide range of power ranges, up to several kilowatts, but the simplest, lowest-cost multi-mode diode is the broad-stripe diode, typically generating up to a few watts of output power.

Standard RE-doped fibres cannot be cladding-pumped. Instead, special fibres known as double-clad fibres are used. These comprise a secondary, multi-mode waveguide into which the multi-mode pump beam can be launched. Typically, this structure is realized with a secondary cladding ('outer cladding') surrounding the original cladding ('inner cladding'). The outer cladding has a lower refractive index than the inner cladding, so that the pump beam can be guided within the inner cladding via total internal reflection. There is still a core for guiding signal light. It resides in the inner cladding and is similar to the core of a core-pumped EDFA, i.e. it is doped with erbium (or another RE) and has a higher refractive index than the surrounding inner cladding (figure B5.10). A single-mode core is preferred for amplifiers in optical transmission systems. The pump light still reaches the RE-doped core and excites the RE ions, so that gain is created and the signal can be amplified.

Thus, a cladding-pumped fibre amplifier has a potential not only for higher power, but also lower cost, than a core-pumped one. However, there are challenges with this technology. While the pump has a high power, its intensity (power density) is much lower than it is with core pumping. This is a problem since erbium ions require a minimum pump intensity to reach gain. A related problem is that because the overlap between the pump beam and the RE-doped region is small, the pump absorption is low. Since the intrinsic absorption of Er^{3+} is already quite low, and since the maximum Er^{3+} concentration is relatively limited, this can lead to excessively long Er-doped fibres, e.g. 100 m or even more if special care is not taken. For this reason a small inner cladding is preferred for EDFAs, significantly smaller than 100 μm diameter [39], but this makes it more difficult to launch the pump light into the fibre. Another proposed [40] and recently demonstrated [41] design modification is to dope the erbium in a ring around the core, since this reduces reabsorption and signal-induced gain-compression. The most common approach, however, is to co-dope (sensitize) the erbium with ytterbium [42]. In an ytterbium-sensitized EDFA, pump photons are initially absorbed by Yb^{3+} ions, which are thus excited. Then, the energy is transferred nonradiatively from Yb^{3+} ions to Er^{3+} ions, resulting in de-excitation of the Yb^{3+} ions and excitation of the Er^{3+} ions. Ytterbium has a larger absorption cross-section than erbium, as well as a much broader absorption band, from 910 to 980 nm. Significantly, the broad absorption can alleviate the requirement of pump wavelength stabilization via temperature control of the pump-LD. Furthermore, Yb^{3+} is comparatively resistant to a reduction of efficiency at high Yb concentrations (concentration quenching), and can therefore be incorporated in much higher concentrations than Er^{3+}. In total, this means that a high gain can be obtained with much lower pump intensities than without sensitization, and even in fibres with a large inner cladding, a fully adequate pump absorption of several decibels/metre can be reached. However, so far, efficient Yb-sensitized EDFAs require a host glass with a high phosphorus content. Unfortunately, this leads to a relatively narrow gain spectrum, so Yb-sensitized EDFAs have not been able to cover the whole C-band. This is a very important aspect.

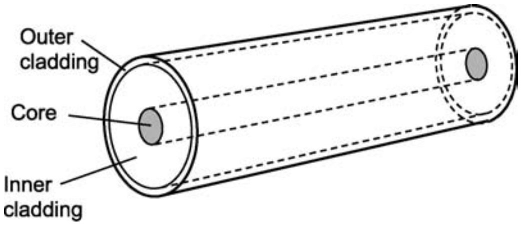


Figure B5.10. Schematic drawing of a double-clad fibre, comprising an RE-doped core, and an inner and an outer cladding.

Ytterbium-free EDFAs do not suffer from this problem—their gain spectrum even defined the extent of the C-band. But, because of the poor pump absorption, Yb-free cladding-pumped EDFAs are relatively inefficient and of low output power. With cladding-pumping, the L-band actually appears more attractive than the C-band, for Yb-sensitized EDFAs because the gain spectrum is wider there, and for Yb-free EDFAs because the pump absorption is higher.

The pump launch is another challenge with cladding-pumped EDFAs. Wavelength-selective fused fibre couplers used to combine the signal and pump beams in core-pumped EDFAs do not work with multi-mode beams. Micro-optic couplers do work and have been used for cladding-pumping. However, these rely on lenses and wavelength-selective reflectors with free-space (unguided) beam propagation to combine the beams, and require precise assembly and alignment. They are much more complex and therefore more costly than fused fibre couplers. Their power-handling capability is also worse, which is a distinct disadvantage for cladding-pumped EDFAs designed for high powers.

Instead of launching the pump through the end of the erbium-doped fibre, together with the signal, one can launch it through the side of the fibre. We will describe three such side-pumping schemes.

One scheme utilizes several (typically three) parallel fibres combined into a common fibre assembly. The fibres are embedded in a polymer coating with a low refractive index so that the fibres themselves can guide pump light. Furthermore, the fibres are in optical contact with each other, meaning that pump light can couple from one fibre into another. One (or several) of the fibres contain a core doped with erbium, typically sensitized with ytterbium. Figure B5.11 shows an example of such a multi-fibre assembly, known as a GTwave fibre. In the GTwave fibre, the fibre with a doped core can be spliced to any fibre carrying a signal. The signal is then launched into the erbium-doped core and amplified in the GTwave fibre. The fibres without a core constitute pump fibres, to which pigtailed pump diodes can be spliced, as many as one to each end of each pump fibre. Once in the pump fibre, the pump light will couple over to the Er-doped fibre and excite the ions in the core.

GTwave fibre devices take simplicity to the limit in that no other components, bar pump-LDs, are required to realize an EDFA (though practical amplifiers do include isolators and, optionally, gain-flattening filters). Another attractive feature is the possibility to include several parallel Er-doped fibres in a multi-port amplifier for parallel amplification. Amplification in eight parallel fibres has been demonstrated [43]. Single-channel GTwave amplifiers can reach well over 1 W of output power with high gain and noise characteristics typical for EDFAs. Broadband versions have been realized both in the C- and L-band [44, 45].

V-groove side-pumping (VSP) is another approach to cladding-pumping, that offers many of the advantages that GTwave fibres provide, with comparable performance [46, 47]. An additional advantage is that VSP amplifiers can be made with any double-clad fibre. On the other hand, while conceptually simple, the fabrication of the v-groove as well as the pump launch are relatively complex issues. A v-groove is fabricated in the side of a double-clad fibre, penetrating into the inner cladding but leaving the (erbium-doped) core intact. Then, pump light from a laser diode is launched through the

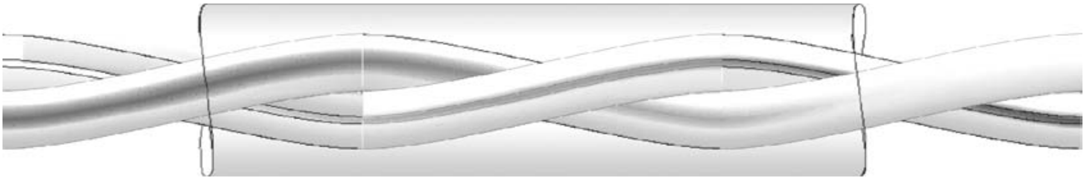


Figure B5.11. A GTwave fibre assembly consisting of a single pump fibre and a single signal fibre, surrounded by a polymer coating. Pump light (red) is launched into the pump fibre, and couples over to the signal fibre. Here, it excites the erbium in the core, leading to amplification of the signal light.

opposite side of the fibre, and hits the v-groove from within the fibre. The light is then deflected off the v-groove facet by ~90°, along the fibre axis and inside the inner cladding, via total internal reflection. Thus, the pump light is launched into the inner cladding of the double-clad fibre, while signal-carrying fibres can be spliced to its ends.

Finally, multimode couplers offer a convenient and widely used way of launching pump light into a double-clad fibre through its side. For example, the tapered fibre bundle is a multi-port version of such a coupler [62].

Despite the success of GTwave, VSP, and multimode couplers, the difficulty of combining high output power, efficient operation, and broadband gain (over the whole C-band) in cladding-pumped EDFAs remains. High-power cladding-pumped ytterbium-doped fibre lasers emitting at 980 nm have therefore been developed as a pump source for EDFAs [48, 49, 61]. This is an alternative way of combining cladding-pumping with EDFAs that is compatible with conventional core-pumped EDFAs, including fused fibre couplers. The EDFA is powered by laser diodes, albeit in an indirect way. Thus, the high-power advantage of cladding-pumping can be realized without compromising amplification bandwidth or reliability. Cladding-pumped ytterbium-doped fibre lasers can generate several watts of output power at 980 nm in a single-mode beam.

B5.6 Planar erbium-doped optical amplifiers

The development of integrated optical amplifiers has been intensified due to the demand for more compact and lower-cost solutions for the DWDM systems, in which the vast majority of erbium-doped fibre-based amplifiers have found commercial use. The commercialization of other planar lightwave components (PLCs) such as AWGs, power splitters and optical switch blocks has resulted in a maturing of the PLC manufacturing with silica-on-silicon being the most widely used technology. The next logical step in the evolution of optical components and subsystems is to combine functions realized in passive PLCs with amplifiers, realizing loss-less or amplifying components of high complexity and functionality that at the same time can be mass produced and offer the overall component size to be reduced significantly compared to the fibre- or bulk optic based counterparts.

In recent years, the development of the planar amplifiers has thus moved from the university and research labs into industry, where it is presently being commercialized in a number of different technologies such as erbium-doped silica-on-silicon, ion-exchange in erbium-doped glass, and as SOAs. At the same time, the fibre-based optical amplifiers are decreasing rapidly in size, the smallest now being less than the size of a credit card. Comparing the technologies, the silica-on-silicon technology potentially has the advantages of the erbium-doped fibre-based amplifier (low noise figure, low polarization dependence, and low gain cross-modulation). In addition, it allows monolithic integration of components known from the passive PLC technology. The main obstacles for the planar erbium-doped amplifier remain the propagation loss and the limited length of the waveguide that requires the erbium concentration to be increased with a factor of 10–100 compared to that in normal erbium-doped fibre. As we describe in section B5.7, the efficiency of the amplifier strongly depends on the erbium concentration and at high concentration it is reduced due to the ion–ion energy transfer that is increasing rapidly with concentration and even more as the erbium ions tend to form clusters at high concentrations. This is known as concentration quenching.

The propagation loss of high-quality waveguides has been reduced to a minimum of 1–3 $\mathrm{dB\,m^{-1}}$, which is still orders of magnitude larger than for fibres, but sufficiently low to allow the use of a long waveguide and hence a relatively low erbium concentration. The fact that the propagation loss is non-negligible, however, means that a planar erbium-doped waveguide amplifier will never be as efficient as the fibre-based amplifiers. On the other hand, the pump laser technology has led to high-power pump lasers at significantly reduced prices. This enables the less efficient planar amplifier technology to be used instead of the fibre-based technology, if it offers other advantages in terms of functionality, performance, manufacturability, cost, or size.

The focused development of the erbium-doped silica-on-silicon technology has led to planar amplifiers with sufficiently high gain and low noise figures to be applicable in commercial long-haul or metro systems. On the picture on figure B5.12, a wafer with four PLC chips, each containing 10 individual waveguide amplifiers, is shown. Each of the amplifiers consists of an erbium-doped waveguide section that is curled up in a spiral to achieve sufficient length. 980/1550 nm combiners on all amplifier inputs and outputs are included for pump multiplexing and removal of excess pump power.

These waveguide amplifiers were produced by a modified plasma-enhanced chemical vapour deposition (PECVD) process, in which the erbium precursors and other co-dopants are introduced directly into the core deposition process. This technique allows detailed control of the glass composition, and enables highly uniform glass layers to be deposited. In contrarst to fabrication processes not involving etching of the waveguide core, such as the ion exchange process, the PECVD-based silica-on-silicon manufacturing process allows the erbium doping to be confined to the waveguide core, where the pump intensity is high. Hence, the erbium ions are easily inverted, leading to a low threshold pump power for amplification. A typical gain spectrum for such a waveguide chip is shown in figure B5.13 for a pump power of 100 mW at 976 nm. For this device, a net gain exceeding 14.5 dB has been achieved over the entire C-band. The performance of such amplifiers typically results in noise figures between 4 and 4.5 dB at low input signal powers.

As for all erbium-doped amplifiers, the performance is strongly dependent on the selected amplifier length, erbium concentration and pump power. For optimized waveguide parameters, gain in excess of 16 dB over the C-band is reached for 100 mW of pump power, and for 200 mW pump power more than 25 dB of gain can be achieved over the C-band, while maintaining the low noise figures. Such gain values will suffice for most applications, and thus represent a viable alternative to the fibre-based amplifiers.

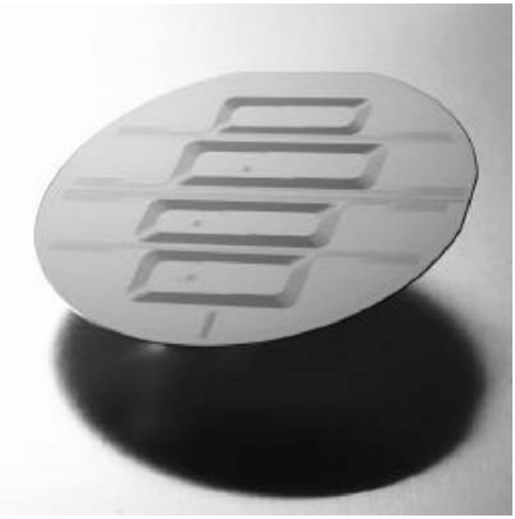


Figure B5.12. Picture of a wafer with four PLC chips, each containing 10 individual waveguide amplifiers. Each amplifier consists of a curled-up erbium-doped waveguide section, and 980/1550 nm combiners on all inputs and outputs for pump multiplexing and removal of excess pump power. The picture is kindly provided by NKT Integration A/S.

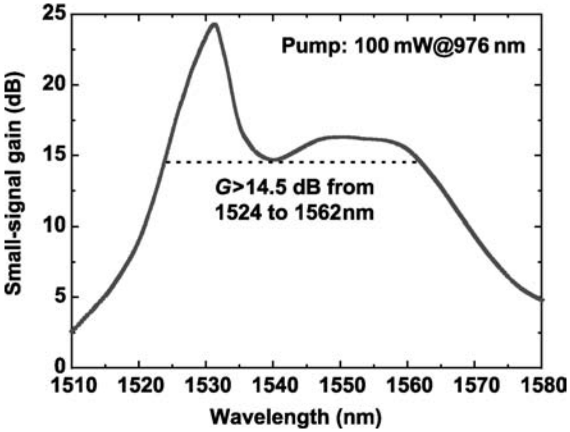


Figure B5.13. A typical gain spectrum for a waveguide chip obtained for a pump power of 100 mW at 976 nm. The results are kindly provided by NKT Integration A/S. Similar gain spectra can be obtained with EDFAs, albeit typically with a higher gain. For WDM applications, the strong gain variations would have to be equalized with a gain-flattening filter.

B5.7 Rare-earth incorporation in silica glass

Practical use of RE-doped silica for amplification purposes typically requires RE concentration levels above 10^{18}–10^{19} cm^{-3}, and elimination of quenching processes. The main sources of quenching are multi-phonon processes associated with O–H bonds and the so-called ‘concentration-quenching’ arising from electrostatic interactions between RE ions. These effects lead to the requirement of a ‘dry’ silica (i.e. with a low hydrogen content) with well-dispersed RE ions. The content of hydrogen in as-fabricated samples can be substantial depending on the fabrication process, but it may be straightforwardly eliminated by subsequent heat treatment and chlorine drying if necessary [17]. Concentration quenching is a more difficult problem to overcome. The root of the problem is that excited RE ions can de-excite each other through electrostatic dipole–dipole interactions, whose magnitudes decay with distance as r^{-6} so that the interaction between neighbouring RE ions in the glass network may be substantial. This problem is particularly severe in the case of erbium, where the energy of the upper laser level ($I_{13/2}$) is close to half the energy of the $I_{9/2}$ level so that a near-resonant upconversion process can take place between close-lying Er ions. Because of its fundamental nature this problem sets an ultimate limit to the useful RE concentration in many systems such as Er- and Nd-doped silica glasses. The problem is magnified by the low solubility of RE ions in the pure SiO_2 matrix, which leads to strong RE clustering during post-production annealing, and in practice it is this effect which limits the usable RE concentrations [18, 19]. This is particularly critical in the case of integrated optics, where the limitations on device dimensions put a lower limit (around 10^{19} cm^{-3}) on the Er concentrations needed to obtain useful signal amplification.

Co-doping of silica glass with aluminium or phosphorus increases the RE solubility and is therefore often used to counter RE cluster formation. Such glasses are known as aluminosilicate and phosphosilicate. Aluminium doping has proven very effective, ensuring essentially complete cluster dissolution at RE concentration levels in the range 10^{19}–10^{20} cm^{-3} (approximately 0.1–1% by weight) when the Al/RE concentration ratio is around 10 or higher. Furthermore, Al doping broadens the emission spectrum of erbium, which is highly desirable. Using P as a co-dopant a slightly higher ratio of about 15 is needed [20]. Nevertheless, with sufficient P content, phosphosilicate glass has a very high RE solubility. This is also true for phosphate glasses (free from silica), in which RE contents of 35% by weight have been realized (however with severe quenching). However, the gain spectrum of phosphosilicate EDFAs is narrower than that of aluminosilicate EDFAs.

At RE concentration levels encountered in Er-doped amplifiers (say, up to 1% by weight), the macroscopic properties of the silica glass are not strongly affected by the co-doping and integration with other glass-based components is a straightforward matter. This suggests that the mechanism underlying the cluster dissolution is a local interaction between RE and co-dopant ions rather than an overall modification of the SiO_2 network structure. Whereas the effect of P codoping is not well understood on a microscopic level there is some experimental [21] and theoretical [22] evidence for this hypothesis in the case of Al codoping. General thermodynamic arguments suggest that cluster dissolution should be favoured by a high fictive temperature of the glass [22], but no systematic experimental data exist to test this hypothesis. In addition, thermodynamic arguments may not apply to glasses fabricated by low-temperature processes such as sol–gel or PECVD. In fact, for Er-doped PECVD glass it is an advantageous strategy to keep the processing temperatures as low as possible in order to limit Er diffusion and thereby cluster formation [19].

An important effect of the local environment around RE dopants is to modify the level spacings and optical matrix elements within the 4f manifold due to ligand-field effects. The site-dependent Stark shifts are an advantage for device fabrication, since they facilitate the manufacturing of amplifiers with wide and flat gain curves suitable for operation at multiple wavelengths. It has been found that the addition of Al leads to an increase of the inhomogeneous broadening by more than a factor of two compared to pure or P-doped silica [23]. The ligand fields also lead to a slight mixing of the f-orbitals with states of even parity, in effect making f–f transitions dipole allowed. As it turns out, this mechanism controls the oscillator strengths of the f-shell transitions. The optical intensities of the f-shell transitions are usually rationalized through the so-called Judd–Ofelt theory [24, 25], which expresses all oscillator strengths in terms of three phenomenological parameters which are related to the matrix elements of the local electrostatic field, but which are most often used as fitting parameters. The Judd–Ofelt theory is approximate, and the underlying assumptions were only partially justified in a recent theoretical analysis of the local fields and electronic structure around an Er impurity [22]. However, in most practical cases the theory is able to describe all oscillator strengths of RE impurities to an accuracy of 10–30% with suitable parameter choices [26–28]. Most systematic investigations of the dependence of Judd–Ofelt parameters on chemical composition of silicate glasses have focused on glass containing significant amounts (>10 mol %) of dopants such as Na, Al, B, Mg or Ca [26, 27, 29, 30]. Even for such mixed glasses the Judd–Ofelt parameters show only slow and moderate (within a factor of 2–3) variation with composition, suggesting that the possibilities for engineering these quantities by varying impurity concentrations is limited in the lightly doped, so-called high-silica, glasses commonly used for fibre amplifiers. Interestingly, the Judd–Ofelt parameters found in Al-codoped fibre preforms [28] differ substantially from those found in heavily doped silicates, possibly due to a more covalent character of the Er–O bonding in the weakly Al-doped silica glass.

B5.8 Summary

EDFAs have revolutionized optical communication systems and are widely deployed. They have removed the loss limit and thereby extended the optical reach to over 10 000 km. They have also enabled mass deployment of WDM systems, and transmission of several terabits per second has been demonstrated in laboratories. The use of lossy components, including optical switches for all-optical networks, is made possible by EDFAs. The only limitation of EDFAs is their limited bandwidth, but they have the potential to cover over 140 nm (~18 THz) in split band configurations and even the more standard 30 nm bandwidth is fully sufficient for systems considered today.

Distributed Raman amplification can provide the best system performance, but deployment is so far very limited. Several amplifier technologies (semiconductor, Raman, OPAs) can operate at any wavelength that is considered for telecoms. Bandwidths of 20 THz or more have been realized in a continuous band. The rapid progress in this field continues towards better optical amplifiers that will find even more uses in the telecom systems of tomorrow. In particular metro applications benefit from the ongoing cost reduction and the development of simple amplifiers [16].

Acknowledgments

We would like to acknowledge the significant help provided by Dr Thomas Feuchter from NKT Integration A/S for his assistance in writing section B5.6, and providing the most recent results on planar erbium-doped optical amplifiers.

References

[1] Agrawal G P 2000 *Nonlinear Fiber Optics* 3rd edn (San Diego, CA: Academic)

[2] Simon J C 1983 Semiconductor laser amplifier for single mode optical fiber communications *J. Opt. Commun.* **4** 51–62

[3] Koester C J and Snitzer E 1964 Amplification in a fiber laser *Appl. Opt.* **3** 1182

[4] Mears R J, Reekie L, Jauncey I M and Payne D N 1987 Low-noise erbium-doped fibre amplifier operating at 1.54 μm *IEE Electron. Lett.* **23** 1026

[5] Agrawal G P 1992 *Fiber-Optic Communication Systems* (New York: Wiley)

[6] Yariv A 1990 Signal-to-noise considerations in fiber links with periodic or distributed optical amplification *Opt. Lett.* **15** 1064–1066

[7] Stubkjaer K E, Mikkelsen B, Durhuus T, Joergensen C G, Joergensen C, Nielsen T N, Fernier B, Doussiere P, Leclerc D and Benoit J 1993 Semiconductor optical amplifiers as linear amplifiers, gates and wavelength converters *ECOC'93: Proc. 19th Eur. Conf. on Optical Communication*, 1:TuC5

[8] Bjarklev A 1993 *Optical Fiber Amplifiers: Design and System Applications* (Boston, MA: Artech)

[9] Hansen S L, Dybdal K and Larsen C C 1992 Upper gain limit in Er-doped fiber amplifiers due to internal Rayleigh backscattering *Optical Fiber Communications, OFC'92* (San Jose, CA) I:TuL4

[10] Mazurczyk V J and Zyskind J L 1993 Polarization hole burning in erbium doped fiber amplifiers *CLEO'93: Conf. on Lasers and Electro-Optics* (Baltimore, MD, May 2–7, 1993) CPD26

[11] Hattori K, Kitagawa T, Oguma M, Ohmori Y and Horiguchi M 1994 Erbium-doped silica-based planar-waveguide amplifier integrated with a 980/1530-nm WDM coupler *Optical Fiber Communications, OFC'94* (San Jose, CA) 1:FB2

[12] Kumar P, Kath W L and Li R-D 1994 Phase-sensitive optical amplifiers *Integrated Photonics Research IPR'94* (San Francisco, CA) 1:SaB1

[13] Bjarklev A 1997 Optical amplifiers *The Communications Handbook* ed J D Gibson (Boca Raton: CRC Press) Chapter 62

[14] Kani J and Jinno M 1999 Wideband and flat-gain optical amplification from 1460 to 1510 nm by serial combination of a thulium-doped fluoride fibre amplifier and fibre Raman amplifier *IEE Electron. Lett.* **35** 1004–1006

[15] Masum-Thomas J, Crippa D and Maroney A 2001 A 70 nm wide S-band amplifier by cascading TDFA and Raman fibre amplifier *OFC'2001* paper WDD9

[16] Clesca B 2002 Optical amplification techniques tussle for metro dominance *Lightwave Europe* pp 22–23

[17] Stone B T and Bray K L 1996 Fluorescence properties of Er^{3+}-doped sol–gel glasses *J. Non-Cryst. Solids* **197** 136–144

[18] Sen S and Stebbins J F 1995 Structural role of Nd^{3+} and Al^{3+} cations in SiO_2 glass: a ^{29}Si MAS-NMR spin–lattice relaxation, ^{27}Al NMR and EPR study *J. Non-Cryst. Solids* **188** 54–62

[19] Sckerl M W, Guldberg-Kjær S, Rysholt Poulsen M, Shi P and Chevallier J 1999 Precipitate coarsening and self-organization in erbium-doped silica *Phys. Rev. B* **59** 13494

[20] Arai K, Namikawa H, Kumata K, Honda T, Ishii Y and Handa T 1986 Aluminium or phosphorus co-doping effects on the fluorescence and structural properties of a neodymium-doped silica glass *J. Appl. Phys.* **59** 3430–3436

[21] Arai K, Yamasaki S, Isoya J and Namikawa H 1996 Electron-spin-echo envelope-modulation study of the distance between Nd^{3+} ions and Al^{3+} ions in the codoped SiO_2 glasses *J. Non-Cryst. Solids* **196** 216–220

[22] Lægsgaard J 2002 Dissolution of rare-earth clusters in SiO_2 by Al codoping: A microscopic model *Phys. Rev. B* **65** 174114

[23] Zemon S, Lambert G, Andrews L J, Miniscalco W J, Hall B T, Wei T and Folweiler R C 1999 Characterization of Er^{3+}-doped glasses by fluorescence line narrowing *J. Appl. Phys.* **69** 6799–6811

[24] Judd B R 1962 Optical absorption intensities of rare-earth ions *Phys. Rev.* **127** 750

[25] Ofelt G S 1962 *J. Chem. Phys.* **37** 511

[26] Takebe H, Morinaga K and Izumitani T 1994 Correlation between radiative transition probabilities of rare-earth ions and composition in oxide glasses *J. Non-Cryst. Solids* **178** 58–63

[27] Li H, Li L, Vienna J D, Qian M, Wang Z, Darb J G and Peeler D K 2000 Neodymium(III) in alumino-borosilicate glasses *J. Non-Cryst. Solids* **278** 35–57

[28] Zemon S, Pedersen B, Lambert G, Miniscalco W J, Andrews L J, Davies R W and Wei T 1991 Excited-state absorption cross sections in the 800-nm band for Er-doped Al/P silica fibers: measurements and amplifier modeling *IEEE Photon. Technol. Lett.* **3** 621–624

[29] Tanabe S and Hanada T 1996 Local structure and 1.5 μm quantum efficiency of erbium doped glasses for optical amplifiers *J. Non-Cryst. Solids* **196** 101–105

[30] Hehlen M P, Cockroft N, Gosnell T R and Bruce A J 1997 Spectroscopic properties of Er^{3+} and Yb^{3+} doped soda-lime silicate and aluminosilicate glasses *Phys. Rev. B* **56** 9302–9318

[31] Siegman A E 1986 *Lasers* (Mill Valley, CA: Univ. Sci. Books)

[32] Genoa corporation *Product information* www.genoa.com

[33] Hansryd J and Andrekson P A 2001 Broad-band continuous-wave-pumped fiber optical parametric amplifier with 49-dB gain and wavelength-conversion efficiency *IEEE Photon. Technol. Lett.* **13** 194–196

[34] Hansryd J, Andrekson P A, Westlund M, Li J and Hedekvist P O 2002 Fiber-based optical parametric amplifiers and their applications *IEEE J. Sel. Top. Quantum Electron.* **8** 506–520

[35] Ho M-C, Uesaka K, Marhic M E, Akasaka Y and Kazovsky L G 2001 200-nm-bandwidth fiber optical amplifier combining parametric and Raman gain *J. Lightwave Technol.* **19** 977–981

[36] Spiekman L H, van den Hoven G N, van Dongen T, Sander-Jochem M J H, Binsma J J M, Wiesenfeld J M, Gnauck A H and Garrett L D 2000 Recent advances in SOA's in WDM applications *Proc. Eur. Conf. on Optical Communication (ECOC 2000)* (München) vol 1 (VDE) pp 35–38

[37] Tangdiongga E, Crijns J J J, Spiekman L H, van den Hoven G N and de Waardt H 2002 Performance analysis of linear optical amplifiers in dynamic WDM systems *IEEE Photon. Technol. Lett.* **14** 1196–1198

[38] Tachibana M, Laming R I, Morkel P R and Payne D N 1991 Erbium-doped fiber amplifier with flattened gain spectrum *IEEE Photon. Technol. Lett.* **3** 118–120

[39] Minelly J D, Chen Z J, Laming R I and Caplen J E 1995 Efficient cladding pumping of an Er^{3+} fibre *Proc. Eur. Conf. on Optical Communication (ECOC 1995)* (Brussels 17–21 Sept. 1995) paper Th.L.1.2, pp 917–290

[40] Nilsson J 1998 Cladding-pumped erbium-doped fiber amplifiers for low-noise high-power WDM and analogue CATV boosters—new design using ring-doping *Optical Fiber Communication Conference, vol 2, OSA Technical Digest Series* (Washington, DC: Optical Society of America) pp 38–39

[41] Bousselet P *et al* 2001 30% power conversion efficiency from a ring-doping all silica octagonal Yb-free double-clad fiber for WDM applications in the C band *Proc. Topical Meeting on Optical Amplifiers and Their Applications (OAA 2001)* post-deadline paper PD1

[42] Minelly J D, Barnes W L, Laming R I, Morkel P R, Townsend J E, Grubb S G and Payne D N 1993 Diode-array pumping of Er^{3+}/Yb^{3+} co-doped fiber lasers and amplifiers *IEEE Photon. Technol. Lett.* **5** 301–303

[43] Alam S U, Nilsson J, Turner P W, Ibsen M, Grudinin A B and Chin A 2000 Low cost multi-port reconfigurable erbium doped cladding pumped fiber amplifier *Eur. Conf. on Optical Communication 2000* (München) paper 5.4.3

[44] Southampton Photonics, Inc. *Product information* www.southamptonphotonics.com

[45] Codemard C, Ylä-Jarkko K, Singleton J, Turner P W, Godfrey I, Alam S-U, Nilsson J, Sahu J K and Grudinin A B 2002 Low noise, intelligent cladding pumped L-band EDFA *Proc. Eur. Conf. on Optical Communication* vol 3 (Copenhagen, Sept. 8–12, 2002) post-deadline paper 1.6

[46] Goldberg L and Koplow J 1998 Compact, side-pumped 25 dBm Er/Yb co-doped double cladding fibre amplifier *Electron. Lett.* **34** 2027–2028

[47] Keopsys, S A *Product information* www.keopsys.com

[48] Zenteno L A, Minelly J D, Liu A, Ellison J G, Crigler S G, Walton D T, Kuksenkov D V and Deineka M J 2001 1 W single-transverse-mode Yb-doped double-clad fibre laser at 978 nm *Electron. Lett.* **37** 819–820

[49] Fu L B, Selvas R, Ibsen M, Sahu J K, Alam S-U, Nilsson J, Richardson D J, Payne D N, Codemard C, Goncharev S, Zalevsky I and Grudinin A B 2002 An 8-channel fibre-DFB laser WDM-transmitter pumped with a single 1.2 W Yb-fiber laser operated at 977 nm *Proc. Eur. Conf. on Optical Communication* vol 3 (Copenhagen, Sept. 8–12, 2002) paper 8.3.5

[50] Goforth D E, Minelly J D, Ellison A J E, Wang D S, Trentelman J P and Nolan D A 2000 *Proc. NFOEC, 2000*, paper B1 (2000)

[51] Arbore M A, Zhou Y, Keaton G and Kane T 2002 36 dB gain in S-band EDFA with distributed ASE suppression *Proc. Topical Meeting on Optical Amplifiers and Their Applications* post-deadline paper PDP4 (Vancouver, July 14–17, 2002)

[52] Mori A and Shimizu M 2003 Ultra-wideband tellurite-based fiber Raman amplifiers *OSA Trends in Optics and Photonics (TOPS) vol 86, Optical Fiber Communication Conference, Technical Digest* (Washington, DC: Optical Society of America) pp 427–429

[53] Stolen R H 1979 Polarization effects in fiber Raman and Brillouin laser *IEEE J. Quantum Electron.* **15** 1157

[54] Emori Y, Kado S and Namiki S 2002 Broadband flat-gain and low-noise Raman amplifiers pumped by wavelength-multiplexed high-power laser diodes *Opt. Fiber Technol.* **8** 107–122

[55] Harwood D W J, Fu A, Taylor E R, Moore R C, West Y D and Payne D N 2000 A 1317 nm neodymium doped fluoride glass waveguide laser *Eur. Conf. on Optical Communication 2000* (München) paper 6.4.8

[56] Mollenauer L F, Grant A R and Mamyshev P V 2002 Time-division multiplexing of pump wavelengths to achieve ultrabroadband, flat, backward-pumped Raman gain *Opt. Lett.* **27** 592–594

[57] Kidorf H, Rottwitt K, Nissov M, Ma M and Rabarijaona E 1999 Pump interactions in a 100-nm bandwidth Raman amplifier *IEEE Photon. Technol. Lett.* **11** 530–532

[58] Session W E 2003 Raman transmission *OSA Trends in Optics and Photonics (TOPS) vol 86 Optical Fiber Communication Conference Technical Digest* (Washington, DC: Optical Society of America) pp 326–336

[59] Codemard C, Soh D B S, Ylä-Jarkko K, Sahu J K, Laroche M and Nilsson J 2003 Cladding-pumped L-band phosphosilicate erbium– ytterbium co-doped fiber amplifier *Topical Meeting on Optical Amplifiers and their Applications* (Otaru, Japan, July 6–9, 2003)

[60] Marhic M 2003 Toward practical fiber optical parametric amplifiers *OSA Trends in Optics and Photonics (TOPS) vol 86 Optical Fiber Communication Conference, Technical Digest* (Washington, DC: Optical Society of America) pp 564–565

[61] Ylä-Jarkko K H, Selvas R, Soh D B S, Sahu J K, Codemard C, Nilsson J, Alam S-U and Grudinin A B 2003 A 3.5 W 977 nm cladding-pumped jacketed-air clad ytterbium-doped fiber laser *ASSP 2003* post-deadline paper PDP2

[62] DiGiovanni D J and Stentz A J 1999 Tapered fiber bundles for coupling light into and out of cladding-pumped fiber devices *US Patent Specification* 5864644

[63] Nilsson J *et al* 2003 Beyond 1 kW with fiber lasers and amplifiers *OSA Trends in Optics and Photonics (TOPS) vol 86, Optical Fiber Communication Conference, Technical Digest* (Washington, DC: Optical Society of America) pp 686–686

[64] Hübner J and Guldberg-Kjær S 2001 Active and passive silica waveguide integration *Proc. ECOC'2001* (Amsterdam, September 2001)

[65] Stolen R H 1980 Nonlinearity in fiber transmission *Proc. IEEE* **69**(10) 1232–1236

[66] Dougherty D J, Kärtner F X, Haus H A and Ippen E P 1995 Measurement of the Raman gain spectrum of optical fibers *Opt. Lett.* **20** 31–33

[67] Lin Q and Agrawal G P 2003 Vector theory of stimulated Raman scattering and its application to fiber-based Raman amplifiers *J. Opt. Soc. Am.* **B20** 1616–1631

Further reading

Siegman A E 1986 *Lasers* (Mill Valley, CA: Univ. Sci. Books)

Agrawal G P 2000 *Nonlinear Fiber Optics* 3rd edn (San Diego, CA: Academic)

Desurvire E 1994 *Erbium-Doped Fiber Amplifiers: Principles and Applications* (New York: Wiley)

Desurvire E, Bayart D, Desthieux B and Bigo S 2002 *Erbium-Doped Fiber Amplifiers: Device and System Developments* (New York: Wiley)

Zyskind J L, Nagel J A and Kidorf H D 1997 Erbium-doped fiber amplifiers for optical communications *Optical Fiber Telecommunications III B*, ed I P Kaminow and T L Koch (San Diego, CA: Academic)

Ellis A and Minelly J D 2002 New materials for optical amplifiers *Optical Fiber Telecommunications IV A—Components*, ed I P Kaminow and T Li (San Diego, CA: Academic)

Srivastava A K and Sun Y 2002 Advances in erbium-doped fibre amplifiers *Optical Fiber Telecommunications IV A—Components*, ed I P Kaminow and T Li (San Diego, CA: Academic)

Anders B 1993 *Optical Fiber Amplifiers: Design and System Applications* (Boston, MA: Artech)

Digonnet M J F (ed) 1993 *Rare Earth Doped Fiber Lasers and Amplifiers* (New York: Dekker)

Sudo S (ed) 1997 *Optical Fiber Amplifiers: Materials, Devices, and Applications* (Norwood, MA: Artech)

France P W (ed) 1991 *Optical Fiber Lasers and Amplifiers* (Glasgow: Blackie–Boca Raton, FL: CRC)

Becker P C, Olsson N A and Simpson J R 1999 *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology* (New York: Academic)

B6
Ultrafast optoelectronics

Günter Steinmeyer

B6.1 Introduction

Optoelectronics is based on electronic devices, which are used for emitting, modulating, transmitting, or sensing light. At a very fundamental stage, these devices require interaction of light with an electronic current, effectively converting photons into electrons or vice versa. The temporal response of an optoelectronic emitter, for example, is therefore always limited by the fastest available rise time of a current-pulse generator. Similarly, detection of light in a photo-detector can only be accomplished directly with a few picosecond temporal resolution [75]. On the electronics side, additional constraints can be imposed by parasitic inductances or capacitances in the electronic circuitry and high-frequency attenuation mechanisms in microwave cables. Streak cameras [71], which merge the generation of photoelectrons and their temporal resolution into one device, overcome some limitations. Nevertheless, even these fastest direct optoelectronic detection devices are typically limited to a response time of the order of 1 ps. The examples discussed so far rely on a direct interaction of light with an electronic current. In the following, we will describe ways to circumvent the electronic bandwidth problem. The fundamental idea behind ultrafast optoelectronics becomes clear from the latest developments for optical communication networks. In early fibre optic data links, light was converted back into an electronic current prior to any processing. This has been replaced by all-optical means of processing photonic data streams. A major improvement in terms of data capacity has been achieved by the method of wavelength-division multiplexing, which allows for terabit/second rates by simultaneously transmitting many channels at different wavelengths through one and the same fibre. In this chapter, we will introduce methods to provide the fundamental optoelectronic functions of emitting, modulating, transmitting, and sensing light, all with a temporal response or resolution of a few femtoseconds. These methods are ultimately limited only by the duration of the optical cycle itself. We will refer to these schemes as *ultrafast optoelectronic devices*, even though the individual optoelectronic components, mediating between photons and electrons, can be inherently slow. The methods described split the optoelectronic process into two steps, an ultrafast all-optical step, which ensures sufficient bandwidth, and a second slow electronic step to allow for efficient conversion between photons and electrons at a strongly reduced bandwidth.

Before continuing our technical discussion in detail, it is appropriate here to indicate how this chapter is organized: first we review methods for the *generation of femtosecond pulses*. Short light pulses can experience strong reshaping effects, which may modify their pulse shape. Compared to electronic pulses, however, these effects start to become significant on terahertz rather than gigahertz bandwidth scales, and they also mainly affect the spectral phase rather than the amplitude. Therefore, particular attention will be paid to methods that allow compensation of *dispersive pulse broadening*. This discussion will be followed by an overview of *femtosecond pulse characterization methods*. Finally, in section B6.5, we will address *methods to modulate phase and amplitude* of femtosecond pulses. Together with the characterization methods, ultrafast optoelectronics nowadays allows for synthesis of desired pulse shapes and control of optical waveforms, very similar to the generation of arbitrary electronic waveforms.

B6.2 Ultrafast laser pulse generation

Compared to electronics, the available bandwidth in optical systems is enormous. This is a clear driving force behind all-optical telecommunications. Fiber-optic systems can provide terabit/second of transmission capacitance over transatlantic distances [2, 76]. From elementary Fourier theorems, it is also clear that the widest bandwidth can be used to support extremely short pulses. The laser material with the widest known gain bandwidth is titanium-doped sapphire [23]. The 650–1100 nm gain bandwidth allows us to generate directly pulses of about 5 fs pulse duration, corresponding to less than two optical cycles of the electric field.

Laser operation is generally sustained by an optical cavity to provide optical feedback into the gain material. The photons circulating in the cavity experience laser gain and losses due to output coupling [84]. Laser gain saturation favours equal filling of the cavity with photons, i.e. cw operation of the laser [57]. In order to generate a short pulse, the energy content of the optical cavity has to be temporally confined into an interval as short as possible. This requires us to introduce a condition which provides an advantage for the laser to operate in short pulses. One means of doing so is to insert an intracavity amplitude modulator, which opens and closes in synchronism with the light travelling through the cavity [85] (figure B6.1). This generates an advantage for those photons which travel during the fully open state of the modulator. If conventional electro-optic or acousto-optic modulators are used, this whole scheme is always limited to some extent by the electronic pulse width of the modulator driver, even though optical pulses shorter than the electronic driving pulses can be generated. Following the philosophy mentioned in our introduction, it is therefore desirable to eliminate electronic bandwidth limitations by switching to an all-optical modulator.

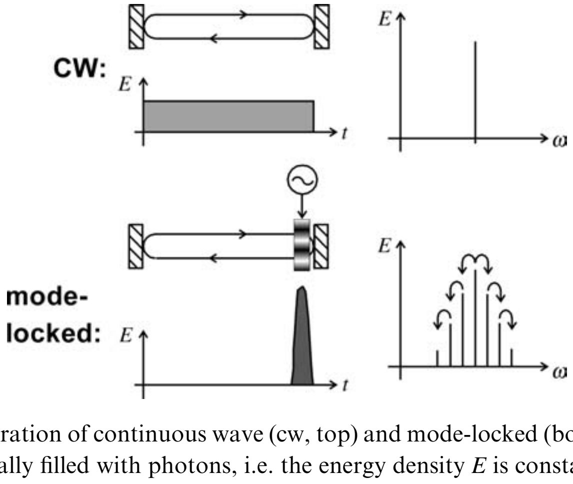


Figure B6.1. Schematic illustration of continuous wave (cw, top) and mode-locked (bottom) operation of a laser. In the cw case, the cavity is equally filled with photons, i.e. the energy density E is constant. Insertion of a modulator synchronously driven at the cavity roundtrip frequency focuses light into a small time slot centred at the fully open position of the modulator (bottom). The two situations are also depicted in the frequency domain on the right side. Continuous operation results in single-longitudinal-mode operation of the laser. The modulator creates sidebands at the neighbouring modes, effectively transferring energy to the spectral wings. This is called mode-locking.

B6.2.1 Saturable absorber mode-locking

The simplest all-optical modulator schemes rely directly on saturable absorption of an optical material [39, 49, 68]. Typically, either organic dyes or semiconductor materials can be used. At high intensities, the absorption of these materials bleaches out, because the majority of electrons or holes has been excited from the ground state to higher energy states. Such *an all-optical modulator* is automatically self-synchronous with the light in the cavity. A small initial power spike inside the cavity will experience less loss than the rest of the cavity's energy content. This positive feedback will self-amplify until all the energy is concentrated in a small time slot. Unfortunately, the relaxation of the bleaching process is not arbitrarily fast, but again, pulses shorter than the modulation time constants can be produced. It is not even necessary for the relaxation time of the absorber to be faster than the cavity round trip. Nevertheless, the all-optical method still experiences a limitation from response-bandwidth effects. Typically, this forbids the generation of pulses much shorter than a picosecond. There are, however, some ingenious schemes that overcome some of the limitations of a slow absorber and extend the operation of saturable absorber mode-locking well into the femtosecond range [28, 38, 52].

Some of these approaches are fairly specialized and rely on the interplay of several optical mechanisms inside the cavity. An alternative approach came from the use of the so-called *reactive nonlinearities* (e.g. the Kerr nonlinearity, see figure B6.2). Other than amplitude changes due to saturable absorption, these effects influence the phase of the light only. This self-phase modulation delays high light intensities in respect to low intensities [1]. As no carrier dynamics are involved in the process, this mechanism can be very fast with a typical response time of less than 1 fs. The 'reactive' character of the nonlinearity calls for a conversion mechanism to transfer the nonlinear phase modulation into an effective amplitude modulation.

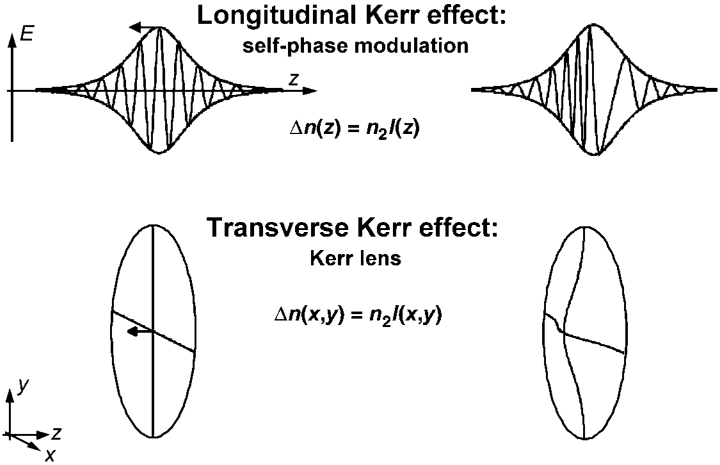


Figure B6.2. The nonlinear optical Kerr effect is caused by a dependence of the index of refraction on intensity. Along the axis of propagation z, this causes a phase retardation of the most intense part of the temporal pulse profile due to self-phase modulation. In the plane perpendicular to z, the retardation causes a deformation of the phase fronts. In the central part of the spatial beam profile, the phase fronts experience an additional curvature, i.e. the Kerr effect causes an effect similar to a lens. Therefore, the transverse Kerr effect is also referred to as a Kerr lens.

B6.2.2 Passive mode-locking based on reactive nonlinearities

We shall now consider the concept of an *effective saturable absorber*. Such a device is actually a combination of a phase-to-amplitude converter and a reactive nonlinearity. Several types of effective saturable absorber have been proposed. One method is to place the self-phase modulator in a second cavity, which is coupled to the gain cavity. The coupled cavity acts as a nonlinear mirror, providing high reflectivity for the more intense centre part of the pulse and low reflectivity everywhere else. This method is called additive-pulse mode-locking [47]. Another method is Kerr-lens mode-locking (KLM) [55, 79, 90]. Here the transverse effect of the Kerr nonlinearity is used instead (see figure B6.2). As the most intense axial region of an optical beam is also phase retarded compared to the outer region, this corresponds to the focusing action of a lens. A suitable arrangement of intracavity apertures (field stops) can then be used as spatial filters to translate the Kerr nonlinearity into an effective absorber. The focused high-intensity light experiences fewer losses at these apertures due to its smaller diameter. Alternatively, one can also arrange the cavity for a better spatial overlap with the pump light when the additional nonlinear lens is in place. Pulses as short as approximately 5 fs have been produced with this method [24, 96].

B6.2.3 Amplification

Typical Kerr-lensed mode-locked lasers can deliver pulse energies of a few nanojoules at a repetition rate of 100 MHz. Pulse energies can be increased either by using a longer cavity [16], by additional extracavity amplification [64], or by cavity dumping [9]. All these methods allow for a few 10 nJ pulse energy at megahertz repetition rates. Further increase of the pulse energy into the microjoule range, whilst still maintaining femtosecond time duration, would increase the peak power of the pulse into a regime where most optical materials are likely to encounter damage. For achieving greater amplification, the pulse therefore needs to be 'stretched' before amplification and then be recompressed to its original duration after the amplifier. The design of the stretcher and compressor will be treated in the following section. The method is called *chirped pulse amplification* (CPA) [6, 94] and has been demonstrated with 15 fs pulses of millijoule energy at a 10 kHz repetition rate [5, 115]. Even stronger amplification and reduction of repetition rate can lead to hundreds of joules pulse energy at 450 fs pulse duration [74]. The latter system reaches a peak power of 1.5 PW (1.5×10^{15} W). The focused intensity reaches $6 \times 10^{20}\,\mathrm{W\,cm^{-2}}$. The electric field strength in the focus of such a laser pulse exceeds typical interatomic binding forces by about three orders of magnitude [12].

B6.2.4 Wavelength conversion—from terahertz to x-rays

So far we have concentrated on direct oscillator and amplifier schemes. Laser materials with wide bandwidth as needed for short-pulse generation are only available in the near-infrared and part of the visible spectral range [17, 27, 28, 118]. Another approach for the generation of short pulses is conversion of femtosecond radiation by nonlinear optical mechanisms. One mechanism is *frequency-doubling* in nonlinear optical crystals. For frequency-doubling of extremely short pulses, typically very thin crystals with a thickness in the few-micrometre range have to be employed [31]. Another method is *optical parametric amplification*, which can be used both in the near-infrared and visible spectral range. The parametric process splits an input photon into two of lower energy, with wavelengths depending on the phase matching conditions in the nonlinear optical material. Beta-barium borate (BBO) has been shown to provide extremely wideband phase-matching, which has been used to generate pulses well below 5 fs [82, 116]. There is a wealth of other methods leading deeper into the UV, including Raman side-band generation [111], high-harmonic generation [12, 58, 61], and Thomson scattering [80]. Wavelength conversion is also a very important mechanism to generate radiation of longer wavelengths than directly available from oscillators. Again, parametric processes can be used here [56]. Another important case is *terahertz radiation* [89]. All these mechanisms allow access to wavelength regions that are not readily covered by wideband laser materials.

B6.3 Femtosecond pulse propagation effects and dispersion compensation

In microwave electronic systems, a severe limitation is imposed by high-frequency damping mechanisms. In optics, limitations due to spectral absorption are typically not a concern or can be easily avoided. Many dielectric media, like glasses and crystals, are transparent in the range of 150–1000 THz [103]. Limitations typically only arise in optical amplification or nonlinear optical conversion. In a 10 THz window in the near infrared (1.55 μm wavelength), exceptionally low losses of 0.3 dB km^{-1} have been demonstrated in silica fibres [70]. In this exact wavelength region, Er-doped glass amplifiers [11, 18] can easily be embedded into optical telecommunication systems. Compared to electronic systems, optical bandwidth is therefore abundant and a much lesser concern.

B6.3.1 Group delay dispersion as leading-order propagation effect

If one tries to launch a 100 fs pulse train into an optical fibre, one finds that the extremely short pulses already broaden to picoseconds after only a few metres of propagation because of dispersion. Dispersion causes different spectral components of the pulse to propagate at different group velocities, which induces broadening of short pulses during propagation. Compensation of dispersive effects therefore poses a ubiquitous problem not only in telecommunication systems [2], but also with ultrashort pulse generation systems [107]. A pulse with angular carrier frequency $\omega = 2\pi c/\lambda$ experiences a *group delay*

$$\mathrm{GD}(\omega) = l\frac{\mathrm{d}}{\mathrm{d}\omega}\frac{\omega}{c}n(\omega) \tag{B6.1}$$

when propagating through a dispersive medium with index n and length l. The group delay determines the propagation time of a pulse and must not be confused with the phase delay $l\,n/c$. To first order, pulse broadening is governed by the *group delay dispersion*

$$\mathrm{GDD}(\omega) = l\frac{\mathrm{d}^2}{\mathrm{d}\omega^2}\frac{\omega}{c}n(\omega). \tag{B6.2}$$

Gaussian pulses of duration τ_0 are stretched to a duration τ, where

$$\tau = \tau_0\sqrt{1+\left[\frac{|\mathrm{GDD}|^2}{\tau_0}\right]^2} \tag{B6.3}$$

when propagating through a material with dispersion GDD. Equation (B6.3) is a useful relation to estimate the severity of pulse broadening in optical systems [1]. Optical materials can exhibit either negative or positive dispersion, and this can change sign as the wavelength varies (see figure B6.3). Fused silica, for example, shows zero dispersion at 1.3 μm with positive dispersion below and negative dispersion above this wavelength. At the *zero-dispersion wavelength* of a material, broadening effects due to GDD are eliminated; but similar effects are caused by higher-order derivatives of the refractive index, e.g. *third-order dispersion*

$$\mathrm{TOD}(\omega) = l\frac{\mathrm{d}^3}{\mathrm{d}\omega^3}\frac{\omega}{c}n(\omega). \tag{B6.4}$$

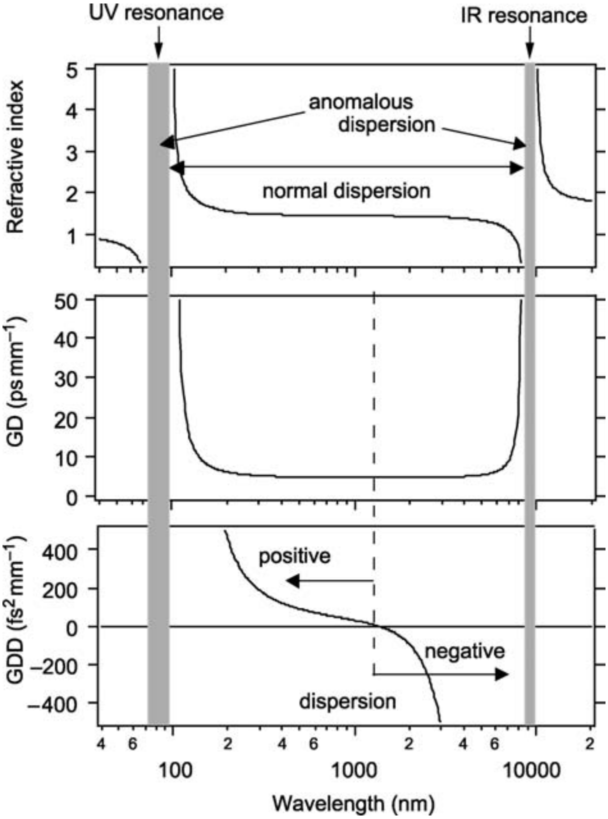


Figure B6.3. Material dispersion. The dispersion of a dielectric material (simplified model of fused silica) with a UV resonance at 80 nm and a vibronic IR resonance at 10 μm is schematically shown. The top figure shows the refractive index itself. For the entire range between the resonances, $\mathrm{d}n/\mathrm{d}\lambda \leq 0$ holds, which is referred to as normal dispersion. Only on resonance, we find anomalous dispersion $\mathrm{d}n/\mathrm{d}\lambda > 0$. The resulting group delay (GD) and group delay dispersion (GDD) are also shown. The GDD is the leading term responsible for pulse reshaping during propagation. Note that regions of positive and negative GDD do not coincide with those of normal and anomalous dispersion.

Positive and negative GDD must not be confused with normal or anomalous first-order dispersion, which refers to the sign of $\mathrm{d}n/\mathrm{d}\omega$. In a system with positive dispersion, blue spectral components will be retarded relative to red components. This causes a variation of the pulse carrier frequency with time, which is usually referred to as a chirp. Depending on the sign of the dispersion, one talks about positive or negative chirp. If no other mechanism is present, both signs of dispersion are totally equivalent in terms of pulse broadening. However, self-phase modulation will typically generate a positive chirp. As self-phase modulation also introduces spectral broadening, its combination with negative dispersion can be used for pulse compression schemes. Balancing of a positive chirp, generated by the nonlinear optical process of self-phase modulation, and negative material dispersion can lead to self-stabilizing optical pulses called *solitons*. These solitons can propagate over great distance through dispersive systems without changing their pulse shape [1, 37].

The discussion so far makes it clear that control and engineering of dispersion is of paramount importance for ultrafast optical systems and for telecommunications. Dispersion management is particularly important for long-distance fibre links [87, 88]. With ever-wider bandwidth becoming accessible, compensation of higher-order dispersion becomes a consideration [73]. Pulse compression is a major mechanism in ultrashort pulse generation [1, 35]. Passively mode-locked lasers (described above in its basic function) make extensive use of recompression of pulses, employing self-phase modulation in the laser crystal together with negative dispersion in the cavity [13, 40]. Only by fully exploiting this mechanism can the shortest pulses be generated. In general, an ultrafast pulse compressor can only be built with negative dispersion, which unfortunately is not a characteristic available for optical materials below 1 μm wavelength. This calls for alternative concepts to compensate for material dispersion and chirps caused by nonlinear optical mechanisms.

One can classify sources of dispersion into bulk or *material dispersion* (i.e. from homogeneous materials like glasses and crystals), *geometrical dispersion* (prism and grating arrangements), *dispersion from interferometric effects*, and *microstructured dispersion* (fibre Bragg gratings, chirped mirrors, chirped quasi-phase matched crystals, arrayed-waveguide gratings). Bulk dispersion has already been treated earlier. Reference data for many materials can be found in [103].

B6.3.2 Geometrical dispersion—prism and grating compressors

In the following, we will firstly address geometrical dispersion, as can be produced by prism [29, 81] and grating sequences [100]. When a short pulse is sent into a prism or on to a grating, its spectral components are angularly dispersed and sent into different directions (see figure B6.4). A second prism of opposite alignment can then be used to make the spectrally dispersed beams parallel again. On their propagation between both prisms, the outer rays have experienced a delay relative to the centre ones. It is important to note that this 'parabolic' spectral delay is equivalent to negative GDD. It can therefore be used to compensate positive material dispersion. Pairs of Brewster-cut prisms can compensate dispersion without introducing losses and have been very successfully used inside laser cavities [29].

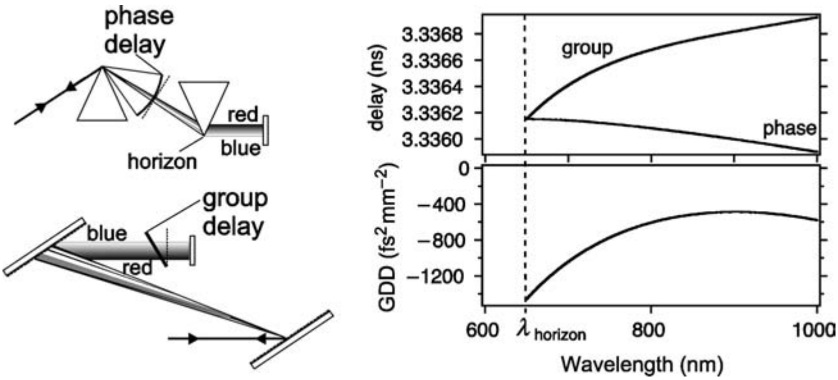


Figure B6.4. Geometric dispersion as caused by prism and grating sequences. Both prisms and gratings exhibit angular dispersion. A beam with broad input spectrum is dispersed into different direction of propagations, with all beams originating at one and the same location at the tip of the first prism. Positions with equal phase delay therefore describe a circle centred at the prism tip. To leading order, the phase delay relative to the centre beam is parabolic. The second prism only serves to render all beams parallel again. Typically, such an arrangement is used in double pass with a retroreflecting mirror as shown. The beam propagating exactly from tip to tip marks the short-wavelength horizon of the prism compressor. A parabolic phase delay front corresponds to a linear group delay (as shown in the grating compressor). An exact calculation for the group and phase delay of a prism compressor (fused silica Brewster prisms with 1 m apex distance) based on [81] is also shown. The resulting GDD is depicted below. Note the strong higher-order dispersion of this approach.

The major shortcoming of the geometrical approach, however, is that it introduces higher-order dispersion terms. For prism compressors, a careful choice of the prism material can give vanishing third-order dispersion in the wavelength range above 800 nm [59]. In particular, fused silica prism pairs introduce vanishing third-order aberrations in the Ti:sapphire wavelength range, which was used for the first demonstration of sub-10 fs pulse generation with this laser [119].

Grating sequences may be used instead for the same purpose (see figure B6.4). They are of extreme importance for CPA [65, 94], which allows for the amplification of pulses from the oscillator up to the millijoule or, at a reduced repetition rate, even to the joule level. To prevent optical damage in the amplifier chains, the oscillator pulse is stretched into the picosecond range before amplification. This reduces its peak power by the stretching ratio and also prevents significant nonlinear optical effects. After amplification, the pulse can then be recompressed into the femtosecond range, using a grating sequence with exactly the opposite dispersion of the stretcher plus any amplifier material dispersion. This restores a short pulse duration and allows for the generation of extremely high peak powers. The *stretcher* [120] employed in CPA is typically a grating sequence, which incorporates a telescope with -1 magnification. The telescope exactly inverts the dispersion of a compressor in all orders. The trick is now to slightly unbalance a stretcher and a matched compressor and to accommodate for material dispersion using the difference between stretcher and compressor dispersion. Second-order dispersion (equation (B6.2)) is adjusted by a difference in grating distances; third-order dispersion (equation (B6.3)) can be zeroed out by adjusting grating angles [60]. Finally, fourth-order aberrations can be compensated by the use of gratings having different line spacings [91]. Typically, aberrations of the telescope have to be compensated by using suitably corrected optics [14]. This approach has been used for the demonstration of 15 fs amplified pulse duration [5, 115]. Other approaches exist, which introduce a controlled amount of imaging aberration in the stretcher's telescope to achieve compensation up to fourth order [95].

B6.3.3 Microstructured dispersion—chirped mirrors and similar devices

One of the major shortcomings of the geometrical dispersion compensation approach is that, with the few exceptions already noted, they typically allow only for the compensation of second-order dispersion. Geometrical dispersion compensation schemes are therefore limited to approximately 100 THz bandwidth. However, the idea of the prism compressor can be readily extended to compensation of arbitrary dispersion: rather than using free-space propagation of laser beams, one could imagine coupling each and every spectral component into an individual fibre of precisely engineered length. A discrete approach would certainly be cumbersome, but integrated optical devices similar in function have been demonstrated and are referred to as *arrayed-waveguide gratings* (AWGs) [22, 86, 98]. The main use of these devices is channel multiplexing in telecom systems; nevertheless their use for dispersion compensation has recently been pointed out [73]. This type of device is shown in figure B6.5, and it is probably the most pictorial example for microstructured dispersion compensation, even though its application is not very widespread yet.

Rather than directly introducing a wavelength dependent propagation length, several other methods for arbitrary dispersion compensation are possible. These are also shown in figure B6.5. One of these approaches is *chirped mirrors* [53, 67, 92, 97]. These dielectric mirrors consist of alternating pairs of transparent high-index and low-index layers. The same effect can be achieved in optical fibres by modifying the refractive index with exposure to UV light through a periodic mask [43, 54]. The portions of the fibre that have been exposed to the short-wavelength radiation show a modified index of refraction. Even though the index differences are much smaller in the fibre grating approach, they provide the same functionality as a distributed Bragg reflector if the period of the index modulation is chirped along the fibre [25, 72]. A Bragg mirror reflects light when all Fresnel reflections at

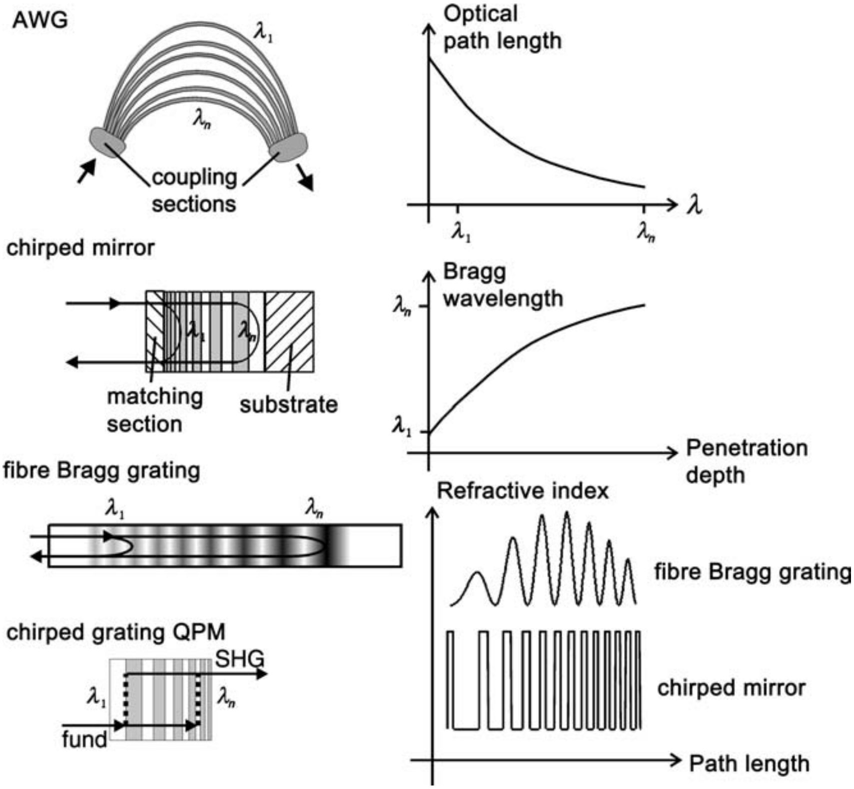

AWG
λ_1
λ_n
coupling sections
Optical path length
λ
λ_1
λ_n
chirped mirror
λ_1
λ_n
matching section
substrate
Bragg wavelength
λ_n
λ_1
Penetration depth
fibre Bragg grating
λ_1
λ_n
Refractive index
fibre Bragg grating
chirped mirror
chirped grating QPM
SHG
λ_1
λ_n
fund
Path length

Figure B6.5. Microstructured dispersion compensation. Shown are four different concepts that can compensate for arbitrarily shaped dispersion. Top: arrayed wave guides (AWGs). The coupling sections provide wavelength-dependent coupling into wave guides of different lengths. This makes the path length a designable function of wavelength. Second: chirped mirrors. Here the Bragg wavelength is varied over the mirror stack, making the penetration depth a function of wavelength. Impedance matching sections are required to reduce detrimental interferometric effects. Third: fibre Bragg gratings. They work similarly to chirped mirrors, but achieve impedance matching by a different apodization method, as shown on the right. Bottom: chirped grating quasi-phase matching. Here the input light is converted into the second harmonic in a quasi-phase-matched crystal. The pooling period of the crystal determines the conversion wavelength, which again allows control of the total group delay as a function of wavelength.

the high/low-index interfaces constructively interfere. This is the case when the optical thickness of all layers is chosen to be equal to a quarter of the light wavelength. Varying the optical layer thickness along the mirror structure then results in a dependence of the wavelength of peak reflectivity, the so-called Bragg wavelength λ_B, on the penetration depth z into the mirror structure. Chirping the mirror structure therefore allows the generation of a desired group delay GD(ω). It is obvious that the Bragg wavelength does not have to be varied linearly with penetration depth; in fact, any single-valued function can be tailored into the Bragg-wavelength chirp function.

It needs to be mentioned that this simple picture is often distorted by other contributions to the dispersion. In chirped mirrors, the top reflection at the interface to air gives rise to undesired interferences, which spoil the dispersion characteristics of the mirror and give rise to strong spectral fluctuations of the dispersion. A solution to this problem is impedance matching from the ambient medium to the mirror stack. In deposited mirror structures, a partial solution can be provided by *double-chirping* [53, 67]. Other methods have been proposed to overcome these dispersion fluctuations [66, 99]. In fibre optics, the UV exposure level can be slowly reduced to give a slow increase of the index modulation in the initial part of the fibre Bragg grating. This is referred to as *apodization* [3].

A novel approach to compensation of arbitrary dispersion is also offered by the method of *quasi-phase matching* (see figure B6.5). Here, the wavelength of conversion of a broadband nonlinear optical conversion process is varied with propagation distance. This means that short wavelengths are converted, e.g. into the second harmonic, at a different propagation distance compared to longer wavelengths. This offers a means to tailor processes like second-harmonic generation in order to support extremely broad bandwidths [32, 46].

B6.3.4 Interferometric effects—Gires–Tournois interferometers

The dispersion of chirped mirrors arises from change in Bragg wavelength of alternating high- and low-index layers. An alternative approach to achieve dispersion compensation is the use of a *Gires–Tournois interferometer* (GTI) [35]. Such an interferometer consists of a partial reflector and a high reflector. This combination reflects all incoming light and has a spectrally flat reflection amplitude response. Its phase response, however, exhibits resonances spaced by $\Delta\nu = c/2L$, similar to a Fabry–Perot interferometer. The spectral phase is a periodic function with regions of negative and positive dispersion. The GTI can be implemented from two air-spaced discrete components, but also using a monolithic mirror structure with a relatively thick spacer layer between quarter wave sections for the partial and high reflector. These structures have been successfully used in femtosecond oscillators [36, 42]. Compared to chirped mirrors, they typically exhibit a lower bandwidth, but they can provide larger values of negative GDD. Their manufacture is not quite so demanding as for chirped mirrors and they can also exhibit very high values of reflectivity [36]. This concept is therefore interesting for lasers having much greater intracavity dispersion and working at longer pulse durations than cavities with chirped mirrors.

B6.4 Measurement of optical waveforms with femtosecond resolution

B6.4.1 Autocorrelation

The major problem in monitoring or characterizing optical waveforms lies in the fact that optical pulses are among the shortest man-made events and there is no shorter controllable event that could be used to sample the waveform. Because of this fundamental limitation, all early characterization methods employed *autocorrelation* [19], a method which characterizes a laser pulse using one and the same pulse, firstly as the sample pulse and secondly as the pulse to be sampled. Using one replica of the input pulse as the reference sample, a second replica is delayed relative to the reference and then optically mixed (i.e. multiplied) with the reference pulse (figure B6.6). Technically, the multiplication of the two optical signals is done using a nonlinear optical effect such as *second-harmonic generation* (SHG) or *two-photon absorption* [78]. As a result, the autocorrelation trace

$$\mathrm{AC}(\delta t) = \int_{-\infty}^{\infty} I(t) I(t - \delta t)\,\mathrm{d}t \tag{B6.5}$$

is measured as a function of time, where $I(t) = E(t)E^{*}(t)$ is the optical intensity [41, 48]. This type of autocorrelation is called *background free*, as it will measure zero signal for large delays $\delta t \to \pm\infty$. A background-free autocorrelator uses a *noncollinear beam geometry*, in such a way that SHG requires one photon from each of the two beams while SHG from each individual beam is not phase matched. This background-free set-up allows for large dynamic ranges, but is typically not the preferred set-up in

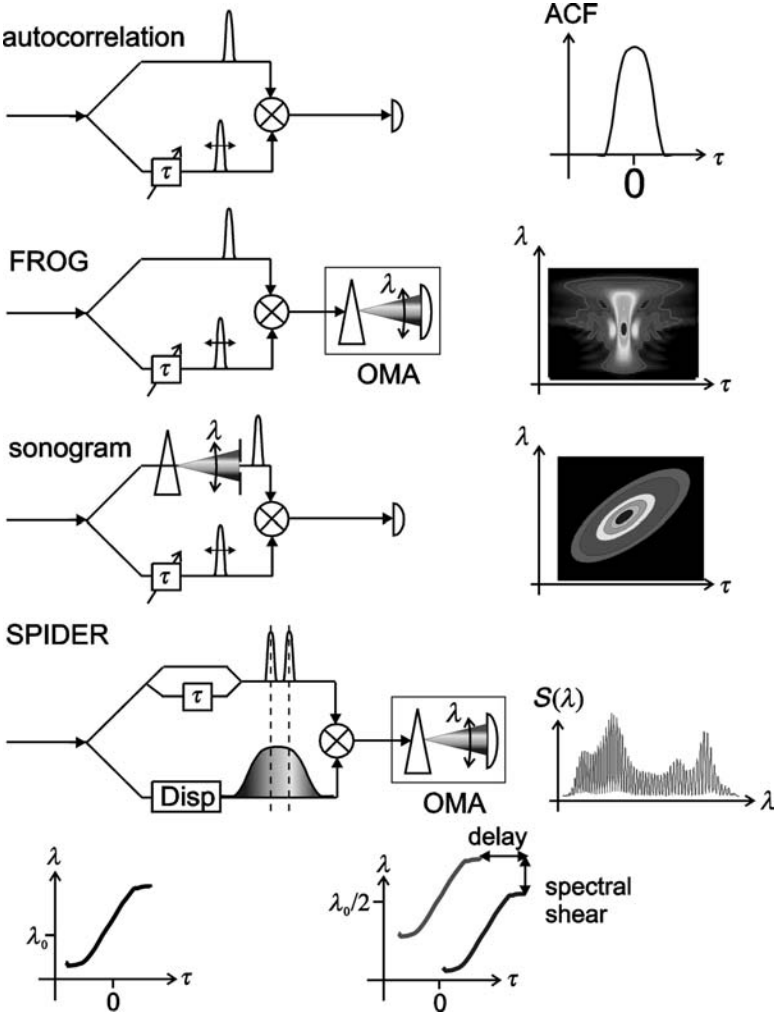


Figure B6.6. Pulse characterization methods. From top to bottom: autocorrelation. The input pulse is split into two identical replicas. One of them serves as the reference sample and is temporarily delayed relative to the other. Multiplication of both replicas in a process such as second-harmonic generation or two-photon absorption delivers the autocorrelation function versus delay time shown on the right. This allows a coarse estimation of the pulse duration. Frequency-resolved optical gating (FROG) additionally spectrally disperses the autocorrelation function and delivers a wavelength-resolved autocorrelation function called a FROG trace. Other than simple autocorrelation, FROG allows complete reconstruction of the input pulse profile. Related is the sonogram technique, which spectrally resolves one of the input replicas instead. Spectral phase interferometry for direct electric-field reconstruction (SPIDER) also creates two replicas of the input pulse at a fixed delay and mixes these two with a chirped copy of the input pulse. The resulting two upconverted replicas are spectrally sheared with respect to each other. The resulting spectral interference pattern $S(\lambda)$ allows reconstruction of the spectral phase of the input pulse.

the sub-10 fs regime. In a *noncollinear set-up*, an additional problem occurs due to '*beam smearing*' caused by the finite crossing angle of the two beams [8]. Therefore a *collinear set-up* is preferred. This then yields the *interferometric autocorrelation* trace [20]:

$$\mathrm{IAC}(\delta t) = \int_{-\infty}^{\infty} \left|\left[E(t) + E(t - \delta t)\right]^2\right|^2 \mathrm{d}t. \tag{B6.6}$$

Unfortunately, no way exists to retrieve the original pulse profile from any measured autocorrelation traces without additional knowledge. Being inspired by an expected theoretical description of the mode-locking process, one can sometimes assume an expected pulse shape, and then retrieval is simple. This is, unfortunately, not a valid assumption in the sub-10 fs regime with its complex pulse shapes. In this regime, simple analytical functions can no longer be assumed for *decorrelation* of the measured autocorrelation function. Additionally, the sub-10 fs regime is very demanding, and pulse shaping by spectral filtering or dispersion in the beam splitters and nonlinear crystal have to be kept to a minimum. Wherever possible, this regime calls for the use of metal-coated reflective optics.

Methods have been discussed to solve the problem of decorrelation [9, 69]. Decorrelation methods require additional experimental information, which, in the simplest form, can be provided by a simultaneous measurement of the power spectrum of the laser. Decorrelation methods employ a computer optimization strategy to find a simultaneous fit to the measured spectrum and autocorrelation. This removes the arbitrariness of assuming a particular pulse shape for pulse retrieval, but, for reliable operation, requires data with excellent signal-to-noise ratio. A practical example based on decorrelation of the IAC trace (equation (B6.6)) and the spectrum is shown in figure B6.7.

B6.4.2 Frequency-resolved optical gating and sonogram

We shall now describe advanced methods called *frequency-resolved optical gating* (FROG) and sonogram to determine the pulse shape. One can conceptually understand the FROG method [101, 102] and the *sonogram technique* [15, 77] as a further extension of the decorrelation methods. For FROG, the autocorrelation of equation (B6.5) is spectrally resolved for each and every delay step (see figure B6.6). The autocorrelation is then sampled on a $\{\delta t, \omega\}$ grid. The autocorrelation spectrogram of the electric field E of the pulse

$$I_{\mathrm{FROG}}^{\mathrm{SHG}}(\delta t, \omega) = \left|\int_{-\infty}^{+\infty} E(t)E(t - \delta t)\exp(-\mathrm{i}\omega t)\mathrm{d}t\right|^2 \tag{B6.7}$$

is called a *FROG trace*. The sonogram technique is very similar in principle, but cuts out a spectral slice of one of the two replica pulses using a narrowband filter function $F(\Omega - \omega)$ centred at a frequency offset Ω from the pulse centre frequency. This filtered replica is then crosscorrelated with the other unfiltered replica yielding the *sonogram trace*

$$I_{\mathrm{sonogram}}(\delta t, \omega) = \left|\int_{-\infty}^{+\infty} E(\Omega)F(\Omega - \omega)\exp(-\mathrm{i}\Omega t)\mathrm{d}\Omega\right|^2. \tag{B6.8}$$

Both these techniques, FROG and sonogram, record data on a two-dimensional array, rather than recording two one-dimensional data traces as in the decorrelation. Examples of such measurements are shown in figure B6.7. The excess data give rise to an increased robustness of the two-dimensional methods, which result in an improved immunity towards noise. Moreover, FROG and sonogram provide built-in consistency checks (marginals), which allow one to detect experimental flaws, e.g. due to limited phase-matching bandwidth or spectral filtering in the set-up. Beyond SHG-FROG, which can be

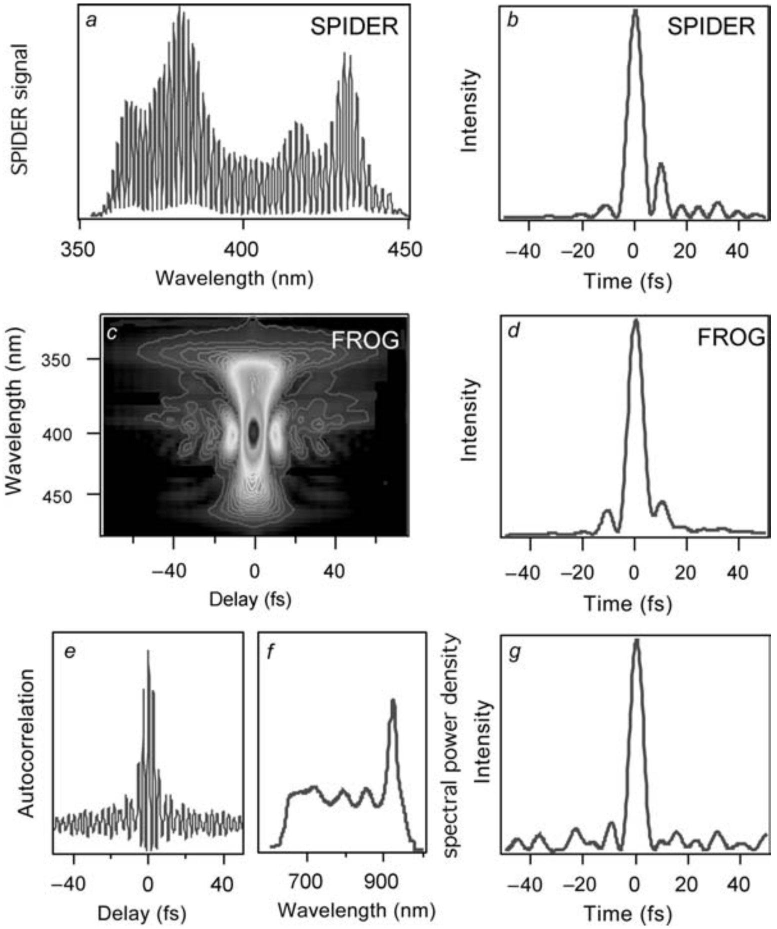


Figure B6.7. Examples of measurements of pulses from a Ti:sapphire laser. Top left: SPIDER measurement and reconstructed pulse [34]. Below: FROG trace and reconstructed pulse measured under nearly identical conditions at the same laser [33]. Bottom part: iteratively reconstructed pulse shape from interferometric autocorrelation and power spectrum of the laser [96]. Note that all three measurements gave compatible pulse durations of about 6 fs or slightly below.

simply understood as an extension of autocorrelation methods, a wide variety of FROG methods have been described. Most notable is *self-diffraction FROG (SD FROG)*, which is very important for measurements on amplifier systems [51]. Here the FROG trace is given by

$$I_{\mathrm{FROG}}^{\mathrm{SD}}(\delta t, \omega) = \left| \int_{-\infty}^{+\infty} E(t)^2 E(t - \delta t)^* \exp(-\mathrm{i}\omega t)\mathrm{d}t \right|^2. \tag{B6.9}$$

The main advantage of SD FROG is that it cannot suffer from limited phase-matching bandwidths as the SHG variant does and can be used in the deep UV. SD FROG generates a sinusoidal intensity pattern in a dielectric medium by crossing two beams under a small angle. This pattern creates an index grating via the Kerr effect and causes self-diffraction of both beams. One of the self-diffracted beams is then spectrally resolved and detected, similar to the SHG variant. This FROG variant, however, requires significantly higher pulse energies and cannot be used for the oscillator-level pulse energies. FROG can also be used in a cross-correlation variant to characterize one unknown pulse with the aid of another known one. This method is called XFROG and is of particular interest for the characterization of pulses in the UV and infrared spectral range [62, 63]. This variant of FROG has been demonstrated with the extremely complex pulse shapes of white-light continua generated in microstructure fibres [112].

Similar to the one-dimensional methods, FROG and the sonogram technique use an optimization strategy for pulse retrieval. Mathematically, it can be strictly shown that knowledge of the FROG trace of a pulse unambiguously defines the pulse's amplitude and phase except for *time reversal*, i.e. one does not know what is the front and what is the back of the pulse. This ambiguity can be removed with an additional measurement. In contrast to the one-dimensional case, a solution always exists, even though it can be a time-consuming computation to find this solution. Recent improvements of the FROG technique have led to very sophisticated retrieval procedures, which can rapidly retrieve the pulse from the FROG trace and allow update rates up to several hertz [50].

B6.4.3 Spectral phase interferometry for direct electric field reconstruction

All techniques described so far involve auto- or crosscorrelation, together with spectral resolution to remove any ambiguity in the pulse reconstruction. Recently, a completely different technique based on *spectral interferometry* [30] emerged for the characterization of femtosecond pulses. This technique is called *spectral phase interferometry for direct electric-field reconstruction* (SPIDER) [45]. The spectrum $S(\omega)$ of two identical pulses $I(t)$ with respective temporal delay ΔT is the spectrum of the single pulse $\tilde{I}(\omega)$, multiplied by a spectrally oscillating term. Measuring the spectral fringe spacing $\Delta\omega = 2\pi/\Delta T$ of $S(\omega)$ allows determination of the temporal spacing ΔT of the two pulses. If these two pulses are identical, the fringe spacing of $S(\omega)$ is also strictly constant over the entire spectrum. A spectral interferogram between a chirped and an unchirped pulse, however, allows not only determination of the delay between the pulses, but also the difference in chirp between the two pulses (figure B6.6). This is the fundamental idea of spectral interferometry. SPIDER generates two delayed replicas of the pulse to be measured. It also generates a third pulse from the input pulse. The third is strongly chirped by sending it through a grating sequence, or through a highly dispersive glass block. The dispersion creates a group delay between the red and blue Fourier components of the third pulse (equation (B6.1)). This chirped pulse is then used to frequency-shift the two replicas of the input pulse using sum-frequency generation. Because of their temporal delay and the strong chirp on the upconverter pulse, both replicas are shifted in frequency by different amounts. Measuring the spectral interferogram of the upconverted replicas allows sampling their relative phase delay as a function of frequency. This is now exactly the information needed to reconstruct the spectral phase. Together with an independent measurement of the amplitude spectrum, this yields a complete description of the pulse. Again, this technique has been demonstrated using pulses from a Ti:sapphire laser (see figure B6.7).

The SPIDER method has been demonstrated with sub-6 fs pulses from Ti:sapphire lasers, with compressed pulses from an amplifier system [21] and with optical parametric amplifiers [116]. One of the major advantages of SPIDER is that it offers a direct reconstruction of the pulse profile, rather than requiring computationally intensive optimization strategies. Typically, the acquisition speed is only limited by the read-out speed of the CCD array used in the spectrograph. Acquisition and reconstruction rates of up to 20 Hz have been demonstrated [83], which makes SPIDER an ideal online tool for aligning complex femtosecond laser systems. SPIDER can also be used in combination with pulse shapers [10]. Other than evolutionary strategies, which try to compensate phase distortions of a pulse by optimizing, e.g. its second-harmonic efficiency, SPIDER provides enough information to set directly the spectral phase to generate a bandwidth-limited pulse.

Given the much more concise data of the SPIDER method, this should allow for a direct and much more rapid phase adaption than lengthy evolutionary optimization strategies. The rapid data acquisition capabilities of the SPIDER method can also be exploited in another way to measure spatially resolved temporal pulse profiles. SPIDER can be adapted to spatially resolve measurements using an imaging spectrograph together with a two-dimensional array. This set-up spatially resolves temporal pulse profiles along the axis defined by the entrance slit of the spectrograph. Rather than the methods discussed so far which integrate over the spatial beam profile, the spatial resolution enables building of an ultrafast camera. Such an ultrafast camera can detect differences in pulse width between beam centre and off-centre regions.

Several methods have been introduced to carry the functionality of a simple photo-diode well into the femtosecond range. The most concise diagnosis is offered by FROG and SPIDER, which set a new standard for the determination of pulse parameters. It needs to be emphasized that the additional information provided by either FROG or SPIDER can directly help to improve pulse generation schemes. Baltuska *et al* designed the mirrors, which were used in the generation of some of the shortest pulses generated to date, according to earlier measurements of the uncompressed pulse [7]. Generally, it depends on the experimental constraints whether to prefer FROG, sonogram, or SPIDER for pulse characterization. All these methods are very well adapted to the particular challenges of femtosecond pulse characterization.

B6.5 Phase and amplitude modulation of short optical pulses

Grating or prism sequences, such as those introduced in section B6.3.2, can also be employed to adjust dispersion in an adaptive way. A very common set-up is the so-called *4f zero dispersion delay line* shown in figure B6.8 [108, 109]. This set-up is very similar to the previously described stretcher, but operates at exactly −1 magnification equivalent to an effective grating distance of zero. A first grating disperses the input pulse, then a lens at distance *f* from the grating creates a spectrally dispersed picture of the input. A second identical lens–grating system reimages the spectrally dispersed picture back onto one point at a distance *f* from the second grating. Provided proper alignment, this set-up is totally neutral in terms of dispersion, i.e. the shape of a pulse propagating through a 4*f* assembly is not modified. However, as the pulse is spectrally dispersed in the Fourier plane in the centre of the set-up, a phase or amplitude modulator array at this location allows manipulation of the waveform.

B6.5.1 Liquid crystal arrays

Several approaches exist for the technical implementation of the modulator array. These can be categorized into phase, amplitude and combined amplitude/phase modulators. Another important aspect is pixelation and the number of pixels or degrees of freedom of such a device. Historically, *liquid crystal arrays* were first used in a 4*f* shaper [109]. Liquid crystals can be used for *phase modulators*, and in combination with polarizers they may also serve as *amplitude modulators*. Combined devices consist of two liquid crystal arrays and polarizers and allow the control of both amplitude and phase. Most devices are pixelated with pixel numbers between 128 [110] and 640 [93]. The pixels typically consist of stripes of a few 10 μm width and can be individually electronically addressed. Liquid crystal phase masks have been used for a variety of applications. Used in a phase shaper, they can compensate for arbitrary dispersion. In this regard, the shaper can be understood as a programmable, microstructured dispersive device. It is particularly useful when dispersion is not very well known or where it may change over time. Recently, as an example, an adaptive pulse compressor was used to compress pulses to sub-6 fs pulse duration [113, 114]. Despite the capability to impose an arbitrary spectral modulation on the pulse, one has to emphasize that a phase modulator can never compress pulses beyond the limit imposed by a Fourier transform of the spectral amplitude profile.

B6.5.2 Acousto-optic pulse shaping

The same functionality can be achieved with an *acousto-optic approach*. The acousto-optic device replaces the liquid crystal array and serves both as amplitude and phase modulator. The acoustic wave transmitted into the acousto-optic modulator is generated by an arbitrary waveform generator, synchronized with the kilohertz repetition rate of the laser source. This acoustic wave generates an index pattern in the acousto-optic crystal. This pattern determines the amplitude and phase of spectral components which are diffracted into the first order of the acousto-optic deflector [26, 44, 104]; see figure B6.8. The main application of this technique is with amplified *kilohertz-repetition-rate* pulses for amplitude shaping. A very similar device, the so-called '*dazzler*', was recently introduced [106]. Here, the interaction of acoustic wave and optical wave is collinear rather than perpendicular. The acoustic wave transfers the input pulse entering in one axis of an acousto-optic crystal into the perpendicular axis. Again, an arbitrary waveform allows one to control the location and efficiency for each and every spectral component of the pulse. This allows compensation for the dispersion and spectral narrowing effects in amplifiers [106]. Note that both acousto-optic approaches are suited only for kilohertz repetition rates.

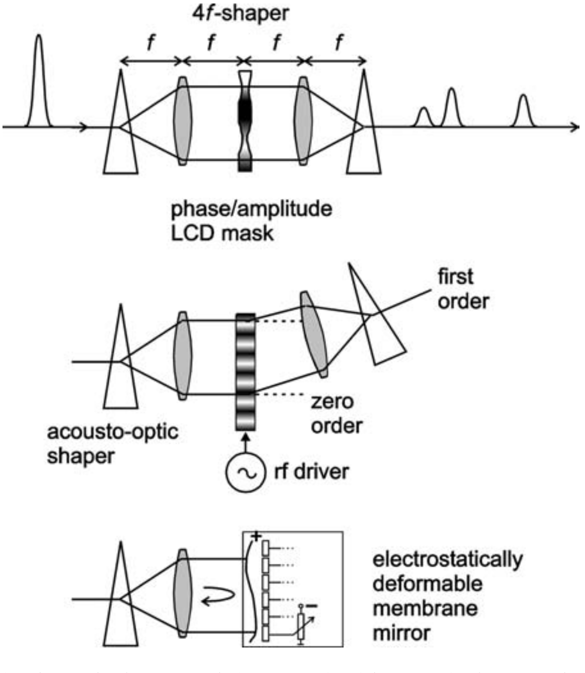


Figure B6.8. Ultrafast amplitude and phase modulators. Liquid crystal phase and amplitude shaper. The input pulse is spectrally dispersed and imaged onto the array with an adjustable retardation/absorption profile. This allows the generation of controllable pulse sequences from an input pulse. The same function can also be achieved by an acousto-optic device. An ultrasonic wave is written into an acousto-optic deflector, where control of the acoustic waveform allows control of the deflection of the optical Fourier components into the first order of the device. Phase control can also be achieved with micro-machined mirror membranes, which can be electrostatically controlled.

B6.5.3 Flexible membrane mirrors

A further approach to phase shaping is to place a mirror with adjustable shape in the Fourier plane of the 4*f* set-up. Such *flexible membrane mirrors* [105] consist of a thin metal-coated silicon nitride membrane. The membrane is suspended over a number of planar electrodes, which act as *electrostatic actuators* (figure B6.8). The membrane is electrically grounded, and a voltage is applied to the actuators, which locally deforms the membrane shape. Typically, these devices have around 10–20 electrodes, and deformations of a few microns can be achieved. For applications in dispersion compensation this complexity is normally more than sufficient. Pulses as short as 7 fs have been generated with this approach [4, 10, 117].

B6.6 Summary

In this chapter, we have highlighted methods to generate and measure ultrafast optical pulses, thereby circumventing the fundamental electronic bandwidth limitations in conventional optoelectronics. The methods described allow for the generation and the characterization of optical pulses much shorter than a picosecond. We have examined propagation effects in optical media and their compensation. Finally, we also addressed set-ups that allow for a manipulation of amplitude and phase of an optical pulse with femtosecond temporal resolution. It is important to note that all these methods are based on rather conventional and relatively slow optoelectronic building blocks, such as cw laser pump diodes for light generation, photodiodes, photomultipliers, or CCD cameras for photodetection, and liquid crystal arrays for the modulation of optical short pulses. The trick lies in exploiting physical optics either by using an ultrafast nonlinear optical effect or by wavelength-division multiplexing. This allows for ultrafast control, without having to control or measure electronic currents on a femtosecond time scale. Telecommunications is a paradigm for the possibilities enabled by optical multiplexing. Transferring more and more of the functionality of a communication network into photonic rather than electronic components has the potential for greatly increased capacities in the near future. Elementary ultrafast optoelectronic building blocks have been introduced in this chapter. These sub-systems open up an avenue for many more applications of photonics in the ultrafast realm.

References

[1] Agrawal G P 1989 *Nonlinear Fiber Optics* (New York: Academic)
[2] Agrawal G P 1997 *Fiber-Optic Communication Systems* (New York: Wiley–Interscience)
[3] Albert J, Hill K O, Malo B, Theriault S, Bilodeau F, Johnson D C and Erickson L E 1995 Apodization of the spectral response of fiber Bragg gratings using a phase mask with variable diffraction efficiency *Electron. Lett.* **31** 222–223
[4] Armstrong M R, Plachta P, Ponomarev E A and Miller R J D 2001 Versatile 7-fs optical parametric pulse generation and compression by use of adaptive optics *Opt. Lett.* **26** 1152–1154
[5] Backus S, Bartels R, Thompson S, Dollinger R, Kapteyn H C and Murnane M M 2001 High-efficiency, single-stage 7-kHz high-average-power ultrafast laser system *Opt. Lett.* **26** 465–467
[6] Backus S, Durfee C G III, Murnane M M and Kapteyn H C 1998 High power ultrafast lasers *Rev. Sci. Instrum.* **69** 1207–1223
[7] Baltuska A, Pshenichnikov M S and Wiersma D A 1998 Amplitude and phase characterization of 4.5-fs pulses by frequency-resolved optical gating *Opt. Lett.* **23** 1474–1476
[8] Baltuska A, Pshenichnikov M S and Wiersma D A 1999 Second-harmonic generation frequency-resolved optical gating in the single-cycle regime *IEEE J. Quantum Electron.* **35** 459–478
[9] Baltuska A, Wei Z, Pshenichnikov M S, Wiersma D A and Szipöcs R 1997 All-solid-state cavity dumped sub-5-fs laser *Appl. Phys. B* **65** 175–188
[10] Baum P, Lochbrunner S, Gallmann L, Steinmeyer G, Keller U and Riedle E 2002 Real-time characterization and optimal phase control of tunable visible pulses with a flexible compressor *Appl. Phys. B* **74** S219–S224
[11] Becker P C, Olsson N A and Simpson J R 1999 *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology* (New York: Academic)
[12] Brabec T and Krausz F 2000 Intense few-cycle laser fields: frontiers of nonlinear optics *Rev. Mod. Phys.* **72** 545–591

[13] Brabec T, Spielmann C and Krausz F 1991 Mode locking in solitary lasers *Opt. Lett.* **16** 1961–1963

[14] Cheriaux G, Rousseau P, Salin F, Chambaret J P, Walker B and Dimauro L F 1996 Aberration-free stretcher design for ultrashort-pulse amplification *Opt. Lett.* **21** 414–416

[15] Chilla J L A and Martinez O E 1991 Direct determination of the amplitude and the phase of femtosecond light pulses *Opt. Lett.* **16** 39–41

[16] Cho S H, Kärtner F X, Morgner U, Ippen E P, Fujimoto J G, Cunningham J E and Knox W H 2001 Generation of 90-nJ pulses with a 4-MHz repetition-rate Kerr-lens mode-locked Ti:Al_2O_3 laser operating with net positive and negative intracavity dispersion *Opt. Lett.* **26** 560–562

[17] Chudoba C, Fujimoto J G, Ippen E P, Haus H A, Morgner U, Kärtner F X, Scheuer V, Angelow G and Tschudi T 2001 All-solid-state Cr:forsterite laser generating 14-fs pulses at 1.3 μm *Opt. Lett.* **26** 292–294

[18] Desurvire E 1994 *Erbium-Doped Fiber Amplifiers: Principles and Applications* (New York: Wiley–Interscience)

[19] Diels J-C and Rudolph W 1996 *Ultrashort Laser Pulse Phenomena* (San Diego: Academic)

[20] Diels J-C, Fontaine J J, McMichael I C and Simoni F 1985 Control and measurement of ultrashort pulse shapes (in amplitude and phase) with femtosecond accuracy *Appl. Opt.* **24** 1270–1282

[21] Dorrer C, de Beauvoir B, Le Blanc C, Ranc S, Rousseau J P, Chambaret J P and Salin F 1999 Single-shot real-time characterization of chirped pulse amplification systems using spectral phase interferometry for direct electric-field reconstruction *Opt. Lett.* **24** 1644–1646

[22] Dragone C 1991 An N × N optical multiplexer using a planar arrangement of 2 star couplers *IEEE Photon. Technol. Lett.* **3** 812–815

[23] Eggleston J M, DeShazer L G and Kangas K W 1988 Characteristics and kinetics of laser-pumped Ti:sapphire oscillators *IEEE J. Quantum Electron.* **24** 1009–1015

[24] Ell R *et al* 2001 Generation of 5-fs pulses and octave-spanning spectra directly from a Ti:sapphire laser *Opt. Lett.* **26** 373–375

[25] Farries M C, Sugden K, Reid D C J, Bennion I, Molony A and Goodwin M J 1994 Very broad reflection bandwidth (44 nm) chirped fibre gratings and narrow bandpass-filters produced by the use of an amplitude mask *Electron. Lett.* **30** 891–892

[26] Fetterman M R, Goswami D, Keusters D, Yang W, Rhee J-K and Warren W S 1998 Ultrafast pulse shaping: amplification and characterization *Opt. Express* **3** 366–375

[27] Fork R L, Cruz C H B, Becker P C and Shank C V 1987 Compression of optical pulses to six femtoseconds by using cubic phase compensation *Opt. Lett.* **12** 483–485

[28] Fork R L, Greene B I and Shank C V 1981 Generation of optical pulses shorter than 0.1 ps by colliding pulse modelocking *Appl. Phys. Lett.* **38** 617–619

[29] Fork R L, Martinez O E and Gordon J P 1984 Negative dispersion using pairs of prisms *Opt. Lett.* **9** 150–152

[30] Froehly C, Lacourt A and Vienot J C 1973 Notions de réponse impulsionelle et de fonction de transfert temporelles des pupilles optiques, justifications expérimentales et applications *Nouv. Rev. Optique* **4** 183–196

[31] Fürbach A, Le T, Spielmann C and Krausz F 2000 Generation of 8-fs pulses at 390 nm *Appl. Phys. B* **70** S37–S40

[32] Gallmann L, Steinmeyer G, Keller U, Imeshev G, Fejer M M and Meyn J-P 2001 Generation of sub-6-fs blue pulses by frequency doubling with quasi-phase-matching gratings *Opt. Lett.* **26** 614–616

[33] Gallmann L, Sutter D H, Matuschek N, Steinmeyer G and Keller U 2000 Techniques for the characterization of sub-10-fs optical pulses: a comparison *Appl. Phys. B* **70** S67–S75

[34] Gallmann L, Sutter D H, Matuschek N, Steinmeyer G, Keller U, Iaconis C and Walmsley I A 1999 Characterization of sub-6-fs optical pulses with spectral phase interferometry for direct electric-field reconstruction *Opt. Lett.* **24** 1314–1316

[35] Gires F and Tournois P 1964 Interféromètre utilisable pour la compression d'impulsions lumineuses modulées en fréquence *C. R. Acad. Sci. Paris* **258** 6112–6115

[36] Golubovic B, Austin R R, Steiner-Shephard M K, Reed M K, Diddams S A, Jones D J and VanEngen A G 2000 Double Gires–Tournois interferometer negative dispersion mirror for use in tunable mode-locked lasers *Opt. Lett.* **25** 275–277

[37] Hasegawa A 1989 *Optical Solitons in Fibers* (Berlin: Springer)

[38] Haus H A 1975 Theory of mode locking with a slow saturable absorber *IEEE J. Quantum Electron.* **11** 736–746

[39] Haus H A 1975 Theory of modelocking with a fast saturable absorber *J. Appl. Phys.* **46** 3049–3058

[40] Haus H A, Fujimoto J G and Ippen E P 1991 Structures for additive pulse modelocking *J. Opt. Soc. Am. B* **8** 2068–2076

[41] Haus H A, Shank C V and Ippen E P 1975 Shape of passively mode-locked laser pulses *Opt. Commun.* **15** 29

[42] Heppner J and Kuhl J 1985 Intracavity chirp compensation in a colliding pulse mode-locked laser using thin-film interferometers *Appl. Phys. Lett.* **47** 453

[43] Hill K O and Meltz G 1997 Fiber Bragg grating technology fundamentals and overview *IEEE J. Lightwave Technol.* **15** 1263–1276

[44] Hillegas C W, Tull J X, Goswami D, Strickland D and Warren W S 1994 Femtosecond laser pulse shaping by use of microsecond radio-frequency pulses *Opt. Lett.* 737–739

[45] Iaconis C and Walmsley I A 1998 Spectral phase interferometry for direct electric field reconstruction of ultrashort optical pulses *Opt. Lett.* **23** 792–794

[46] Imeshev G, Arbore M A, Fejer M M, Galvanauskas A, Fermann M and Harter D 2000 Ultrashort-pulse second-harmonic generation with longitudinally nonuniform quasi-phase-matching gratings: pulse compression and shaping *J. Opt. Soc. Am. B* **17** 304–318

[47] Ippen E P, Haus H A and Liu L Y 1989 Additive pulse modelocking *J. Opt. Soc. Am. B* **6** 1736–1745

[48] Ippen E P and Shank C V 1975 Dynamic spectrometry and subpicosecond pulse compression *Appl. Phys. Lett.* **27** 488

[49] Ippen E P, Shank C V and Dienes A 1972 Passive modelocking of the cw dye laser *Appl. Phys. Lett.* **21** 348–350

[50] Kane D J 1999 Recent progress toward real-time measurement of ultrashort laser pulses *IEEE J. Quantum Electron.* **35** 421–431

[51] Kane D J and Trebino R 1993 Characterization of arbitrary femtosecond pulses using frequency-resolved optical gating *IEEE J. Quantum Electron.* **29** 571–578

[52] Kärtner F X, Jung I D and Keller U 1996 Soliton modelocking with saturable absorbers *IEEE J. Sel. Top. Quantum Electron.* **2** 540–556

[53] Kärtner F X, Matuschek N, Schibli T, Keller U, Haus H A, Heine C, Morf R, Scheuer V, Tilsch M and Tschudi T 1997 Design and fabrication of double-chirped mirrors *Opt. Lett.* **22** 831–833

[54] Kashyap R 1999 *Fiber Bragg Gratings* (New York: Academic)

[55] Keller U, ’tHooft G W, Knox W H and Cunningham J E 1991 Femtosecond pulses from a continuously self-starting passively mode-locked Ti:sapphire laser *Opt. Lett.* **16** 1022–1024

[56] Kobayashi T and Shirakawa A 2000 Tunable visible and near-infrared pulse generator in a 5 fs regime *Appl. Phys. B* **70** S239–S246

[57] Krausz F, Brabec T and Spielmann C 1991 Self-starting passive modelocking *Opt. Lett.* **16** 235–237

[58] L’Huillier A, Lompre L-A, Mainfray G and Manus C 1992 *High-Order Harmonic Generation in Rare Gases*, ed A L’Huillier, L-A Lompre, G Mainfray and C Manus (New York: Academic) pp 139–206

[59] Lemoff B E and Barty C P J 1993 Cubic-phase-free dispersion compensation in solid-state ultrashort-pulse lasers *Opt. Lett.* **18** 57–59

[60] Lemoff B E and Barty C P J 1993 Quintic-phase-limited, spatially uniform expansion and recompression of ultrashort optical pulses *Opt. Lett.* **18** 1651–1653

[61] Lewenstein M, Balcou P, Ivanov M Y, L’Huillier A and Corkum P B 1994 Theory of high-harmonic generation by low-frequency laser fields *Phys. Rev. A* **49** 2117–2132

[62] Linden S, Giessen H and Kuhl J 1998 XFROG—a new method for amplitude and phase characterization of weak ultrashort pulses *Phys. Status Solidi. B* **206** 119–124

[63] Linden S, Kuhl J and Giessen H 1999 Amplitude and phase characterization of weak blue ultrashort pulses by downconversion *Opt. Lett.* **24** 569–571

[64] Liu Z, Izumida S, Ono S, Ohtake H and Sarukura N 1999 High-repetititon-rate, high-average-power, mode-locked Ti:sapphire laser with an intracavity continuous-wave amplification scheme *Appl. Phys. Lett.* **74** 3622–3623

[65] Maine P, Strickland D, Bado P, Pessot M and Mourou G 1988 Generation of ultrahigh peak power pulses by chirped pulse amplification *IEEE J. Quantum Electron.* **24** 398–403

[66] Matuschek N, Gallmann L, Sutter D H, Steinmeyer G and Keller U 2000 Back-side coated chirped mirror with ultra-smooth broadband dispersion characteristics *Appl. Phys. B* **71** 509–522

[67] Matuschek N, Kärtner F X and Keller U 1999 Analytical design of double-chirped mirrors with custom-tailored dispersion characteristics *IEEE J. Quantum Electron.* **35** 129–137

[68] Mocker H W and Collins R J 1965 Mode competition and self-locking effects in a Q-switched ruby laser *Appl. Phys. Lett.* **7** 270–273

[69] Naganuma K, Mogi K and Yamada H 1989 General method for ultrashort light pulse chirp measurement *IEEE J. Quantum Electron.* **25** 1225–1233

[70] Nagel S R, MacChesney J B and Walker K L 1985 *Optical Fiber Communications* vol 1, ed T Li (Orlando: Academic) chapter 1

[71] Nordlund T M 1991 *Streak Cameras for Time-Domain Fluorescence*, ed T M Nordlund (New York: Plenum)

[72] Ouellette F 1987 Dispersion cancellation using linearly chirped Bragg grating filters in optical wave-guides *Opt. Lett.* **12** 847–849

[73] Parker M C and Walker S D 2001 Multiple-order adaptive dispersion compensation using polynomially-chirped grating devices *Appl. Phys. B* **73** 635

[74] Pennington D M *et al* 2000 Petawatt laser system and experiments *IEEE J. Sel. Top. Quantum Electron.* **6** 676–688

[75] Prein S, Diddams S and Diels J C 1996 Complete characterization of femtosecond pulses using an all-electronic detector *Opt. Commun.* **123** 567–573

[76] Ramaswami R and Sivarajan K 1998 *Optical Networks: A Practical Perspective* (San Mateo: Morgan Kaufmann)

[77] Reid D T 2000 Algorithm for complete and rapid retrieval of ultrashort pulse amplitude and phase from a sonogram *IEEE J. Quantum Electron.* **35** 1584–1589

[78] Reid D T, Padgett M, McGowan C, Sleat W E and Sibbett W 1997 Light-emitting diodes as measurement devices for femtosecond laser pulses *Opt. Lett.* **22** 233–235

[79] Salin F, Squier J and Piché M 1991 Mode locking of Ti:Al_2O_3 lasers and self-focusing: a Gaussian approximation *Opt. Lett.* **16** 1674–1676

[80] Schoenlein R W, Leemans W P, Chin A H, Volfbeyn P, Glover T E, Balling P, Zolotorev M, Kim K-J, Chattopadhyay S and Shank C V 1996 Femtosecond x-ray pulses at 0.4 Å generated by 90° Thomson scattering: a tool for probing the structural dynamics of materials *Science* **274** 236–238

[81] Sherriff R E 1998 Analytic expressions for group-delay dispersion and cubic dispersion in arbitrary prism sequences *J. Opt. Soc. Am. B* **15** 1224–1230

[82] Shirakawa A, Sakane I, Takasaka M and Kobayashi T 1999 Sub-5-fs visible pulse generation by pulse-front-matched noncollinear optical parametric amplification *Appl. Phys. Lett.* **74** 2268–2270

[83] Shuman T M, Anderson M E, Bromage J, Iaconis C, Waxer L and Walmsley I A 1999 Real-time SPIDER: ultrashort pulse characterization at 20 Hz *Opt. Express* **5** 134–143

[84] Siegman A E 1986 *Lasers* (Mill Valley, CA: University Science Books)

[85] Siegman A E and Kuizenga D J 1974 Active mode-coupling phenomena in pulsed and continuous lasers *Optoelectronics* **6** 43–66

[86] Smit M K 1988 New focusing and dispersive planar component based on optical phased array *Electron. Lett.* **24** 385–386

[87] Smith N J, Forysiak W and Doran N J 1996 Reduced Gordon–Haus jitter due to enhanced power solitons in strongly dispersion managed systems *Electron. Lett.* **32** 2085–2086

[88] Smith N J, Knox F M, Doran N J, Blow K J and Bennion I 1996 Enhanced power solitons in optical fibres with periodic dispersion management *Electron. Lett.* **32** 54–55

[89] Smith P R, Auston D H and Nuss M C 1988 Subpicosecond photoconducting dipole antennas *IEEE J. Quantum Electron.* **24** 255

[90] Spence D E, Kean P N and Sibbett W 1991 60-fsec pulse generation from a self-mode-locked Ti:sapphire laser *Opt. Lett.* **16** 42–44

[91] Squier J, Barty C P J, Salin F, LeBlanc C and Kane S 1998 Using mismatched grating pairs in chirped pulse amplification systems *Appl. Opt.* **37** 1638–1641

[92] Stingl A, Spielmann C and Krausz F 1994 Generation of 11-fs pulses from a Ti:sapphire laser without the use of prisms *Opt. Lett.* **19** 204–206

[93] Stobrawa G, Hacker M, Feurer T, Zeidler D, Motzkus M and Reichel F 2001 A new high-resolution femtosecond pulse shaper *Appl. Phys. B* **72** 627–630

[94] Strickland D and Mourou G 1985 Compression of amplified chirped optical pulses *Opt. Commun.* **56** 219–221

[95] Sullivan A and White W E 1995 Phase control for production of high-fidelity optical pulses for chirped-pulse amplification *Opt. Lett.* **20** 192–194

[96] Sutter D H, Steinmeyer G, Gallmann L, Matuschek N, Morier-Genoud F, Keller U, Scheuer V, Angelow G and Tschudi T 1999 Semiconductor saturable-absorber mirror-assisted Kerr-lens mode-locked Ti:sapphire laser producing pulses in the two-cycle regime *Opt. Lett.* **24** 631–633

[97] Szipöcs R, Ferencz K, Spielmann C and Krausz F 1994 Chirped multilayer coatings for broadband dispersion control in femtosecond lasers *Opt. Lett.* **19** 201–203

[98] Takahashi H, Nishi I and Hibino Y 1992 10 GHz spacing optical frequency-division multiplexer based on arrayed waveguide grating *Electron. Lett.* **28** 380–382

[99] Tempea G 2001 Tilted-front-interface chirped mirrors *J. Opt. Soc. Am. B* **18** 1747–1750

[100] Treacy E B 1969 Optical pulse compression with diffraction gratings *IEEE J. Quantum Electron.* **5** 454–458

[101] Trebino R, DeLong K W, Fittinghoff D N, Sweetser J, Krumbügel M A and Richman B 1997 Measuring ultrashort laser pulses in the time–frequency domain using frequency-resolved optical gating *Rev. Sci. Instrum.* **68** 1–19

[102] Trebino R and Kane D J 1993 Using phase retrieval to measure the intensity and phase of ultrashort pulses: frequency-resolved optical gating *J. Opt. Soc. Am. A* **10** 1101–1111

[103] Tropf W J, Thomas M E and Harris T J 1995 *Properties of Crystals and Glasses*, ed W J Tropf, M E Thomas and T J Harris (New York: McGraw-Hill) pp 33.1–33.101

[104] Tull J X, Dugan M A and Warren W S 1997 High resolution acousto-optic shaping of unamplified and amplified femtosecond laser pulses *J. Opt. Soc. Am. B* **14** 2348

[105] Vdovin G V 1995 Spatial light modulator based on the control of the wavefront curvature *Opt. Commun.* **115** 170–178

[106] Verluise F, Laude V, Cheng Z, Spielmann C and Tournois P 2000 Amplitude and phase control of ultrashort pulses by use of an acousto-optic programmable dispersive filter: pulse compression and shaping *Opt. Lett.* **25** 575–577

[107] Walmsley I A, Waxer L and Dorrer C 2001 The role of dispersion in optics *Rev. Sci. Instrum.* **72** 1–29

[108] Weiner A M 2000 Femtosecond pulse shaping using spatial light modulators *Rev. Sci. Instrum.* **71** 1929–1960

[109] Weiner A M, Heritage J P and Kirschner E M 1988 High-resolution femtosecond pulse shaping *J. Opt. Soc. Am. B* **5** 1563–1572

[110] Weiner A M, Leaird D E, Patel J S and Wullert J R 1992 Programmable shaping of femtosecond optical pulses by use of 128-element liquid crystal phase modulator *IEEE J. Quantum Electron.* **28** 908–920

[111] Wittmann M, Nazarkin A and Korn G 2001 Synthesis of periodic femtosecond pulse trains in the ultraviolet by phase-locked Raman sideband generation *Opt. Lett.* **26** 298–300

[112] Xu L, Kimmel M W, O'Shea P, Trebino R, Ranka J K, Windeler R S and Stentz A J 2000 Measuring the intensity and phase of an ultrabroadband continuum *Proc. Ultrafast Phenomena XII* (Charleston, SC) pp 129–131

[113] Xu L, Li L M, Nakagawa N, Morita R and Yamashita M 2000 Application of a spatial light modulator for programmable optical pulse compression to the sub-6-fs regime *IEEE Photon. Technol. Lett.* **12** 1540–1542

[114] Xu L, Nakagawa N, Morita R, Shigekawa H and Yamashita M 2000 Programmable chirp compensation for 6-fs pulse generation with a prism-pair-formed pulse shaper *IEEE J. Quantum Electron.* **36** 893–899

[115] Yamakawa K, Aoyoma M, Matsuoka S, Takuma H, Barty C P J and Fittinghoff D 1998 Generation of 16-fs, 10-TW pulses at a 10-Hz repetition rate with efficient Ti:sapphire amplifiers *Opt. Lett.* **23** 525–527

[116] Zavelani-Rossi M, Cerullo G, Silvestri S D, Gallmann L, Matuschek N, Steinmeyer G, Keller U, Angelow G, Scheuer V and Tschudi T 2001 Pulse compression over a 170-THz bandwidth in the visible by use of only chirped mirrors *Opt. Lett.* **26** 1155–1157

[117] Zeek E, Bartels R, Murnane M M, Kapteyn H C, Backus S and Vdovin G 2000 Adaptive pulse compression for transform-limited 15-fs high-energy pulse generation *Opt. Lett.* **25** 587–589

[118] Zhang Z, Nakagawa T, Torizuka K, Sugaya T and Kobayashi K 1999 Self-starting modelocked Cr^{4+}:YAG laser with a low-loss broadband semiconductor saturable-absorber mirror *Opt. Lett.* **24** 1768–1770

[119] Zhou J, Taft G, Huang C-P, Murnane M M, Kapteyn H C and Christov I P 1994 Pulse evolution in a broad-bandwidth Ti:sapphire laser *Opt. Lett.* **19** 1149–1151

[120] Martinez O E 1987 3000 times grating compressor with positive group velocity dispersion: application to fiber compensation in 1.3–1.6 μm region *IEEE J. Quantum Electron.* **23** 59–65

Further reading

Bartels R, Backus S, Zeek E, Misoguti L, Vdovin G, Christov I P, Murnane M M and Kapteyn H C 2000 Shaped-pulse optimization of coherent emission of high-harmonic soft X-rays *Nature (London)* **406** 164–166

Brixner T, Oehrlein A, Strehle M and Gerber G 2000 Feedback-controlled femtosecond pulse shaping *Appl. Phys. B* **70** S119–S124

Drescher M, Hentschel M, Kienberger R, Tempea G, Spielmann C, Reider G A, Corkum P B and Krausz F 2001 X-ray pulses approaching the attosecond frontier *Science* **291** 1923–1927

Farkas G and Toth C 1992 Proposal for attosecond light pulse generation using laser induced multiple-harmonic conversion processes in rare gases *Phys. Lett. A* **168** 447–450

Kaindl R A, Wurm M, Reimann K, Hamm P, Weiner A M and Woerner M 2000 Generation, shaping, and characterization of intense femtosecond pulses tunable from 3 to 20 μm *J. Opt. Soc. Am. B* **17** 2086–2094

Mittleman D M, Jacobsen R H and Nuss M C 1996 T-ray imaging *IEEE J. Sel. Top. Quantum Electron.* **2** 679–692

Paul P M, Toma E S, Breger P, Mullot G, Augé F, Balcou P, Muller H G and Agostini P 2001 Observation of a train of attosecond pulses from high harmonic generation *Science* **292** 1689–1692

Proctor B and Wise F 1992 Quartz prism sequence for reduction of cubic phase in a modelocked Ti:sapphire laser *Opt. Lett.* **17** 1295–1297

Rabitz H, Vivie-Riedle R d, Motzkus M and Kompa K 2000 Whither the future of controlling quantum phenomena? *Science* **288** 824–828

Schwefel H P 1995 *Evolution and Optimum Seeking* (New York: Wiley)

Spielmann C, Burnett N H, Sartania S, Koppitsch R, Schnurer M, Kan C, Lenzner M, Wobrauschek P and Krausz F 1997 Generation of coherent X-rays in the water window using 5-femtosecond laser pulses *Science* **278** 661–664

Warren W S, Rabitz H and Dahleh M 1993 Coherent control of quantum dynamics: the dream is alive *Science* **259** 1581–1589

Wattelier B, Zou J P, Sauteret C, Migus A, Loiseaux B and Huignard J-P 2000 Dynamic phase masks for focal spot optimization and shaping of cw- and high power laser beams *Proc. SPIE* **4421** 42–46

Zare R N 1998 Laser control of chemical reactions *Science* **279** 1875–1879

B7
Integrated optics

Nikolaus Boos and Christian Lerminiaux

B7.1 Introduction

Integrated optics, which has been a research topic for about 20 years, deals with compact single function devices and the integration of multiple optical functionalities on a single chip or into a single package. Ultimately, this is seen as a way to reduce footprint and cost with respect to conventional bulk and micro-optic components, and also to obtain components with increased performance or even new functionalities.

The fabrication technologies have benefited from the experience in the semiconductor industry. However, there are a few main differences between integrated electronics and integrated optics: routing of photons on a chip needs structures with sizes of the order of millimetres, which results in a lower circuit density with respect to electronics. Also, for a given optical function there may be one preferred material, which currently puts a limitation on the development and deployment of monolithic integrated optical circuits. Hybrid integration combines the best materials on a common platform and represents an intermediate solution to this problem. Both monolithic and hybrid integration approaches will probably coexist depending on application, performance and cost.

At present, integrated optics is the technology of choice for a few functions in the optical network. With the increasing maturity of the fabrication technology and the growth in bandwidth and ongoing standardization of optical networks, integrated optic devices will become more cost effective than discrete components.

This chapter is organized as follows: we will start with a description of waveguides and basic elements for phase and polarization control, waveguide couplers and interferometers.

Fabrication processes and properties of common opto-electronic materials will be sketched in section B7.3, followed by device packaging and the techniques for function and material integration.

Section B7.5 gives a brief overview of optical networks and the functions therein, followed by a selection of recent publications on integrated devices to realize these network elements. We will conclude with all-optical components, which will slowly replace electronics in future generation networks.

B7.2 The integrated optics toolbox—waveguides and basic devices

B7.2.1 Waveguides and requirements

Planar optical waveguides are the keys for the construction of integrated optical circuits. In chapters A1.5 and A2.4 a detailed formalism for the propagation of an electromagnetic wave in dielectric media based on Maxwell's equations and the boundary conditions for the electrical and magnetic fields has been developed for:

- Slab waveguides, in which light is confined in only one direction.
- Rectangular waveguides, in which light is confined in two directions. As an exact analytical solution for the electric and magnetic fields cannot be given in this case, a variety of approximation and numerical methods [1, 2] exist. For example, tightly confined modes are needed for low threshold III–V lasers or merely guiding light efficiently from one point to another point, and they represent the foundation of planar integrated optics. Cross-sections for commonly used rectangular waveguides are shown in figure B7.1.

One basic requirement for large-scale integration is that the overall device losses should be low, which in turn translates into an upper limit for the propagation losses through straight and curved waveguides. Exact values depend on the functionality and final device size, but typically $< 0.1\,\mathrm{dB\,cm^{-1}}$ is needed for silica waveguides. There are essentially four loss contributions:

- Absorption in the material through atomic/molecular transitions in amorphous materials, and through excitation of electrons from the valence to the conduction band (interband absorption), or electrons within the conduction band (free carrier absorption) in semiconductor materials.
- Scattering occurs at defects or material inhomogeneities within the guide, or through scattering at the surfaces through roughness induced in the waveguide fabrication process.
- Radiation losses arise in waveguide bends with constant bend radius and in waveguides with discontinuities in the bend radius (transition losses). Bending losses can be minimized by increasing the bend radius R and confining the mode well by increasing the relative index contrast Δ

  $$\Delta = (n_{\mathrm{core}} - n_{\mathrm{clad}})/n_{\mathrm{core}}$$

  Transition losses can be minimized through a lateral offset between waveguide sections or by varying the bend radius adiabatically.
- Substrate leakage—as will be mentioned in section B7.3.1, silica waveguides are commonly fabricated on a silicon substrate, from which they are optically separated via a buffer layer. If this buffer layer is not thick enough, the high index of silicon gives rise to leakage of light from the core into the substrate. As a rule of thumb, the buffer layer should be at least twice as thick as the core layer.

Other losses that play a role in interconnecting optical building blocks arise from a mismatch of the respective mode fields as well as angular or lateral misalignments (see section B7.4.1).

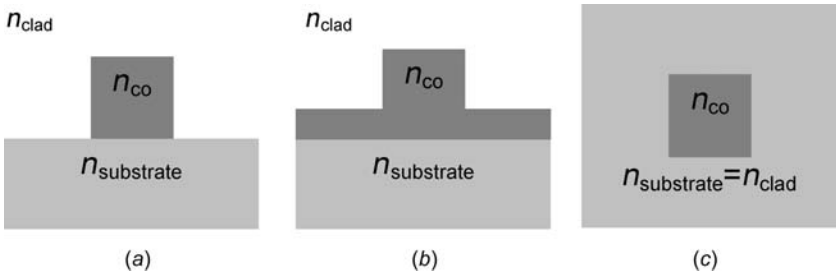


Figure B7.1. Cross-sections through common waveguides. (*a*) The raised strip guide is formed by an etching process removing the higher index material n_{co} on both sides of the guide. (*b*) The rib or ridge waveguide is similar to (*a*), except that the higher index material n_{co} is not completely removed. (*c*) The channel waveguide can be formed by ion implantation, ion exchange/diffusion processes or over-cladding a waveguide as depicted in (*a*).

In the remainder of this section we will give an overview on

- Dynamic index control in waveguides.
- Waveguide couplers and power splitters.
- Interferometric devices—*Mach–Zehnder interferometers*, the *arrayed waveguide grating* (AWG) and *ring resonators*.
- Modelling of integrated functions.

B7.2.2 Dynamic phase and polarization control in optical waveguides

Functional devices often need a dynamic control of the properties of the guided mode (effective index, polarization state), and depending on the application different physical phenomena are used.

The *thermo-optic effect* uses the temperature dependence of the waveguide index $\mathrm{d}n/\mathrm{d}T$ to obtain a phase shift

$$\Delta\phi = \frac{2\pi}{\lambda}\frac{\mathrm{d}n}{\mathrm{d}T}\Delta T \cdot L$$

where L is the length over which the waveguide temperature is changed by ΔT.

In practice, this effect is used in silica ($\mathrm{d}n/\mathrm{d}T = 1\times10^{-5}\,\mathrm{K}^{-1}$ [3]), polymer ($\mathrm{d}n/\mathrm{d}T = -2\cdots-3\times10^{-4}\,\mathrm{K}^{-1}$ [4]), Si ($1.8\times10^{-4}\,\mathrm{K}^{-1}$ [5]) or lithium niobate ($\mathrm{LiNbO_3}$, $\mathrm{d}n/\mathrm{d}T = 5\times10^{-5}\,\mathrm{K}^{-1}$) waveguides for switching or modulating devices with tuning speeds of the order of milliseconds (polymer, silica) to microseconds (silicon).

Electro-optic effects

- Current injection into a semiconductor junction or application of a reverse voltage changes the electron/hole density in the valence and conduction bands of a semiconductor. As a consequence, the absorption for photons with a wavelength close to the band gap energy will change, which in turn results in an index variation, since both parameters are linked through the Kramers–Kronig relations [6, 7].
- Application of an electric field to a dielectric medium leads to an instantaneous induced polarization P [8]

$$\boldsymbol{P} = \varepsilon_0{}^*(\chi(1)\boldsymbol{E} + \chi(2)\boldsymbol{EE} + \chi(3)\boldsymbol{EEE} + \cdots) \equiv P_{\mathrm{Linear}} + P_{\mathrm{Nonlinear}}$$

where $\chi(i)$ are the tensors of the linear ($i = 1$), quadratic ($i = 2$) and higher order susceptibilities. $\chi(1)$ accounts for linear phenomena like absorption and reflection.

The second order susceptibility $\chi(2)$ in materials without inversion symmetry leads to the linear electro-optic (Pockels) effect, and the associated index change is usually expressed as

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1\cdots3} r_{ij}E_j$$

with $j = 1\cdots3$ denoting x-, y- and z-axes and r_{ij} being the electro-optic tensor.

Lithium niobate ($\mathrm{LiNbO_3}$, $r_{33} = 30.9\,\mathrm{pm\,V^{-1}}$ [9]) is widely used for high speed modulation and switching. Induced $\chi(2)$ is possible through poling, which has been reported in silica glasses ($r = 1\,\mathrm{pm\,V^{-1}}$ [10]) and polymer ($r_{33} = 13\,\mathrm{pm\,V^{-1}}$ [11]).

Third order susceptibility $\chi(3)$ allows the control of the refractive index through the electrical field of a second optical signal (all-optical processing). The index change can be parametrized as

$$\Delta n = n_2 I$$

with I being the power density of the optical control signal. In silica, n_2 is of the order of $10^{-16}\,\mathrm{cm}^2\,\mathrm{W}^{-1}$ [10, 8], and therefore is too small for integrated devices. On the other hand, semiconductors exhibit a very large nonlinearity (n_2 of the order of 10^5–$10^6\,\mathrm{cm}^2\,\mathrm{W}^{-1}$) used for switching (section B7.5.4) and all-optical signal processing (section B7.5.6).

We also refer to chapters B1.2, B4 and B6 for a more detailed treatment of electro-optic control.

The *acousto-optic* effect: mechanical strain from a surface or volume acoustic wave induces index changes, which in turn alter the phase of light. Depending on the interaction length between the acoustic wave and the optical mode, one distinguishes between Raman–Nath (short interaction length) and Bragg type modulators.

Magneto-optic control uses the Faraday effect, i.e. linearly polarized light is rotated in the presence of a magnetic field. The Faraday effect can be used for polarization-splitting and nonreciprocal devices such as optical isolators and circulators.

B7.2.3 Waveguide couplers and power splitters

Waveguide couplers and power splitters are needed for the construction of interferometric devices, which will be described in section B7.2.4.

Directional couplers: when waveguides are brought into proximity in such a way that their mode fields partially overlap, power is transferred from one waveguide to another (figure B7.2). For identical lossless waveguides and co-propagating modes, the power coupling ratio is given by (I_0 is the incoming

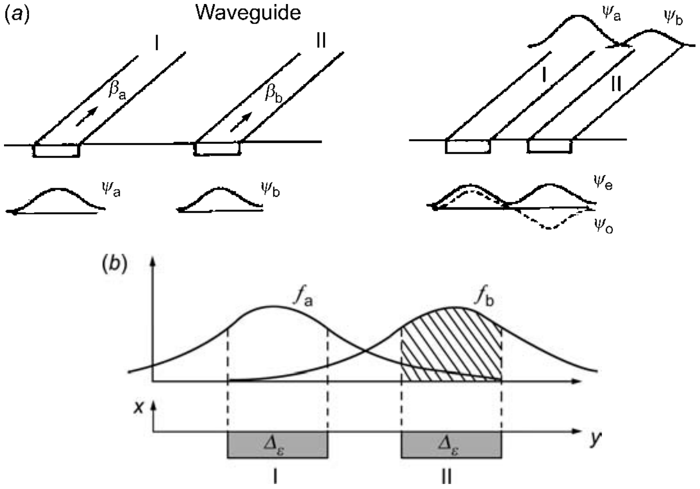


Figure B7.2. Directional coupler [13]. (*a*) Propagating modes in two uncoupled (left) and coupled (right) waveguides. In analogy to a coupled pendulum, the coupled system has both symmetric ψ_e and asymmetric ψ_o fundamental modes (having different propagation constants) which are excited and lead to a periodic energy exchange between modes ψ_a and ψ_b. (*b*) Cross section through a directional coupler. The coupling coefficient κ is proportional to the overlap integral of the shaded areas. Reproduced by permission of The McGraw-Hill Companies.

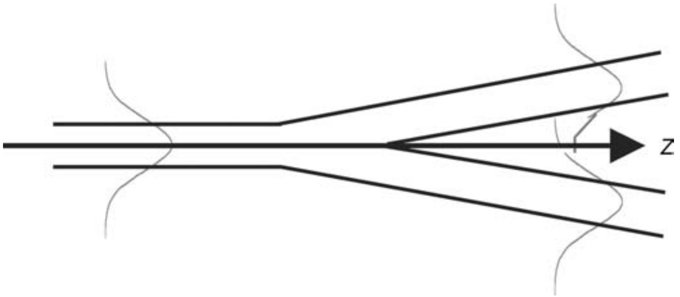


Figure B7.3. Y-splitter—the fundamental mode propagating in the left input waveguide will be equally split between the two output guides. For a symmetric device, the splitting ratio is 50%/50% (3 dB). 2^n splitter trees can be obtained through cascading of Y-splitters.

power, I_1 and I_2 are the powers at the end of the coupler)

$$I_1/I_0 = \sin^2(\kappa z) \equiv \sin^2\left(\frac{\pi}{2L_c}z\right) \quad I_2/I_0 = \cos^2(\kappa z) \equiv \cos^2\left(\frac{\pi}{2L_c}z\right)$$

The coupling coefficient κ (figure B7.2(*b*)) increases with the decreasing gap between the coupled waveguides, and by adjusting the ratio of propagation length z to coupling length L_c (defined as the length needed for complete power transfer), any splitting ratio can be designed. However, the wavelength dependent mode field diameter makes the coupling ratio also wavelength dependent; this can be avoided in more sophisticated designs with tapered asymmetric coupling regions [12]. For a treatment of counter-propagating modes, see reference [13].

On the other hand, the operation of Y-splitters (figure B7.3) depends only on their symmetry and thus the splitting ratio is wavelength independent. Power splitters (1×2^n) can be built by concatenation of Y-splitters.

A simpler way to obtain a large number of input or output ports is the star coupler [14], in which a multimode slab waveguide is placed between a fan of input and output waveguides (figure B7.4).

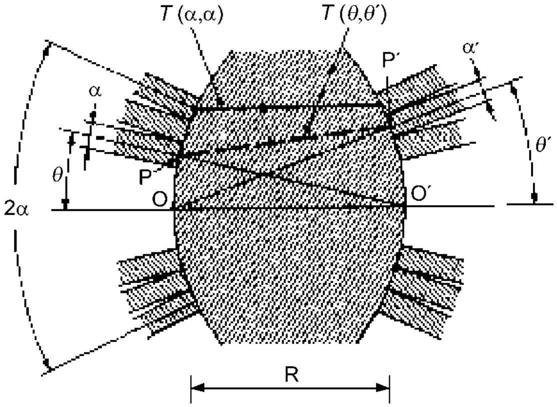


Figure B7.4. Star coupler [14], consisting of an array of input and output waveguides interconnected by a multimode slab guide. Light entering the multimode slab waveguide will diffract and excite fundamental and higher order modes, resulting in mixing and excitation of the guided modes of the output waveguide array. Reproduced by permission of IEEE.

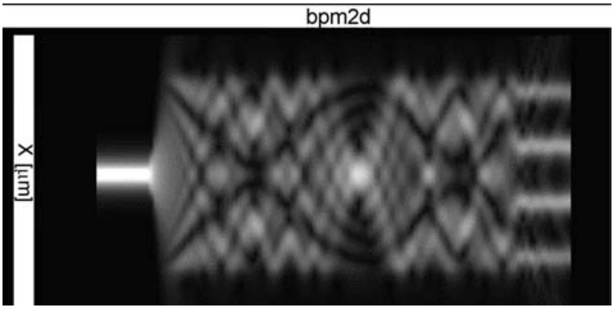


Figure B7.5. Multimode interference MMI coupler [16]. The diagram shows the amplitude distribution in a 1 × 4 splitter as obtained from beam propagation modelling (BPM) [196]. Light entering from the left will excite multiple modes in the middle section, which interfere and create local minima and maxima. Finally, constructive interference is obtained in multiple focal points, and the fundamental mode in the output waveguides is excited.

A uniform splitting ratio is achieved by adequate mutual coupling between input waveguides, and couplers with up to 144 × 144 ports have been published [15].

Multimode interference couplers [16] are based on self-imaging, i.e. the fact that single or multiple images are produced when higher order modes are excited within a wide multimode waveguide, and these modes interfere (constructively or destructively) along the propagation direction of the light. An example of a 1 × 4 splitter is depicted in figure B7.5.

B7.2.4 Interferometric devices—Mach–Zehnder interferometers, the arrayed waveguide grating, ring resonators

The Mach–Zehnder interferometer is a key element for filters, switches and modulators, which will be discussed later in section B7.5. A schematic is shown in figure B7.6—light is split in a first directional

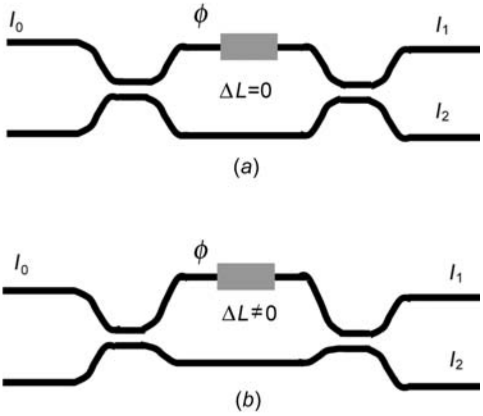


Figure B7.6. Symmetric (*a*) and asymmetric (*b*) Mach–Zehnder interferometer consisting of two directional couplers, two interferometer arms and a dynamic phase shifter. The path from input 0 to output 1 is called 'bar', the path 0 → 2 'cross'.

coupler, passes through the delay lines and then recombines. The transmitted optical power is a function of the phase difference, and is given by

$$I_1/I_0 = \sin^2(\phi/2) \ \text{‘bar’ path}$$

$$I_2/I_0 = \cos^2(\phi/2) \ \text{‘cross’ path}$$

in the case of perfect 50%/50% (3 dB) directional couplers.

For the ‘cross’ path, there will be a constructive interference for $\phi = m{\cdot}2\pi$ and destructive interference for $\phi = (2m+1)\pi$. Tunable devices can be made by dynamic control of ϕ via the effects as described in section B7.2.2.

For broadband devices (i.e. devices having a flat spectral response over a wide passband), such as switches, the phase difference will be within $[0, \pi]$ (‘symmetric MZI’). If spectral features within the passband are desired (as in the case of the equalizing filters described in B7.5.2), the phase difference will rather be chosen to be within $[m{\cdot}2\pi, (m+1)2\pi]$. The fixed part $\phi = m{\cdot}2\pi$ will be generated by a fixed delay of geometrical length in one arm (‘asymmetric MZI’)

$$\Delta L = \frac{m\lambda}{n}$$

and the additional 2π will be kept tunable. The device has a periodic response, which will be repeated for every free spectral range (FSR)

$$\text{FSR} \equiv \Delta\lambda = \frac{n\Delta L}{m(m+1)}$$

The arrayed waveguide grating (AWG) is the analogue to a bulk diffraction grating and can be used in optical networks as a multiplexer/demultiplexer (figure B7.7(*a*)) in wavelength division multiplexed (WDM) transmission systems, or as a wavelength selective building block for add–drop devices or optical crossconnects (OXCs) (sections B7.5 and C1).

The AWG [17, 18] (figure B7.7(*b*)) consists of input/output waveguides, two star couplers and an array of waveguides with a constant path length difference ΔL between adjacent guides. Light coming into the first ‘free propagating region’ (FPR) radiates and excites all waveguides in the grating. After propagation through the array, the light will constructively interfere in one focal point of the second star coupler if the grating condition

$$m\lambda = n\Delta L$$

is fulfilled. n is the effective index of the mode guided in the waveguide array, λ is the wavelength, m is the grating order. The location of the focal point depends on the wavelength as the phase delay between adjacent guides is given by $\Delta L/\lambda$, and the passbands have a Gaussian shape as shown in figure B7.7(*c*).

A ring resonator consists of couplers and a feedback loop of geometrical length L and is the analogue to the free-space Fabry–Perot interferometer.

The two-port resonator shown in figure B7.8(*a*) is the simplest case: it has a unity power transmission and a phase response, making it useful in the synthesis of phase compensating devices [100].

Open ring resonators (figure B7.8(*c*)) have resonances in both amplitude and phase response and may be used in $N \times N$ switch matrices [19] or for bistable devices having a nonlinear medium in the feedback loop [2].

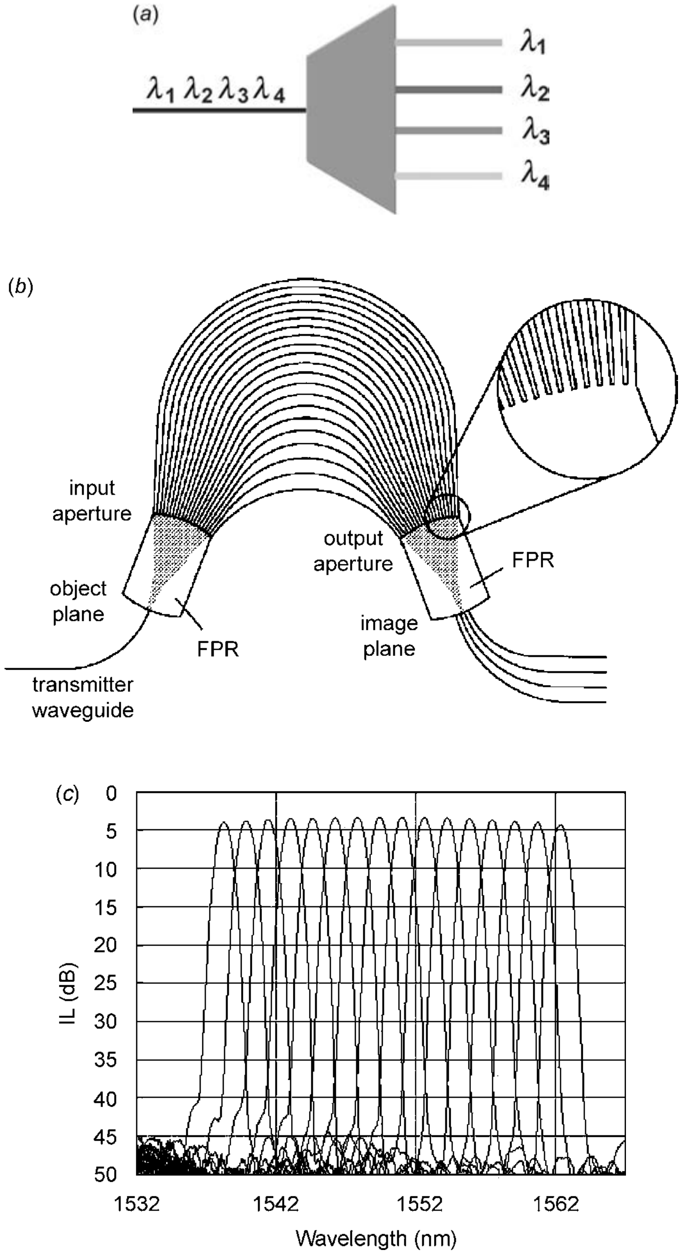

(a)
λ1 λ2 λ3 λ4
λ1
λ2
λ3
λ4
(b)
input aperture
object plane
transmitter waveguide
FPR
output aperture
image plane
FPR
(c)
IL (dB)
0
5
10
15
20
25
30
35
40
45
50
1532
1542
1552
1562
Wavelength (nm)

Figure B7.7. AWG router. (*a*) Basic function. (*b*) Schematic configuration of an AWG [17]: light enters the input star coupler (free propagating region FPR), will then diffract and then excites guided modes in each of the waveguides in the array. In the arrayed waveguide section there is a constant delay ΔL between adjacent arms, and finally constructive interference is obtained in the image plane (second star coupler). (*c*) Spectrum of a 1×16 demultiplexer with Gaussian passband and $<5\,\text{dB}$ fibre–fibre losses [197]. Reproduced by permission of IEEE.

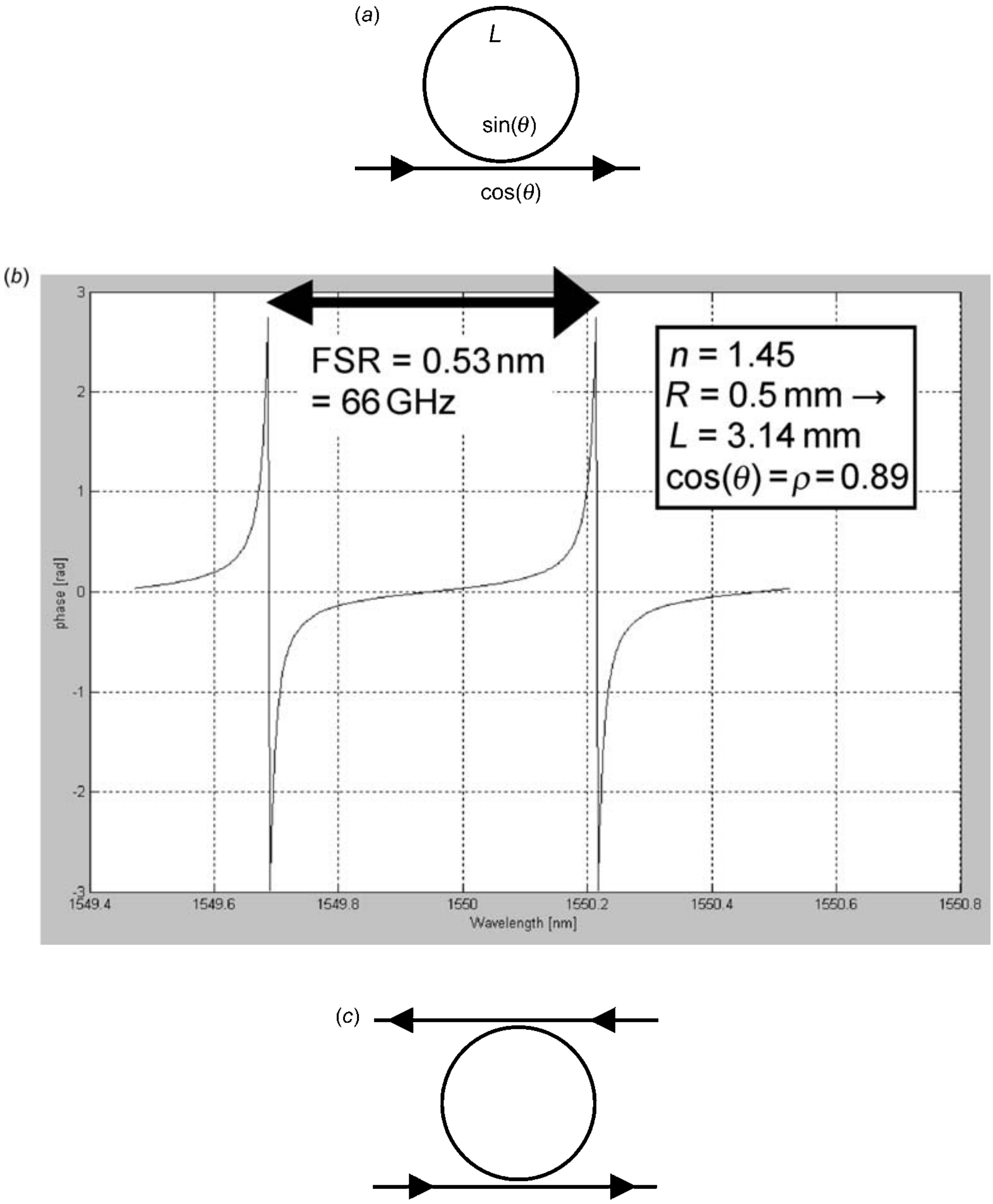


Figure B7.8. Ring resonator. (*a*) Two-port ring resonator consisting of a coupler with coupling angle θ and a feedback loop of length L. (*b*) The phase response calculated from the transfer function given in table B7.1 has a resonance peak which is repeated every free spectral range FSR. (*c*) Open ring resonator.

In analogy to the Mach–Zehnder interferometer, the FSR of a ring resonator is given by

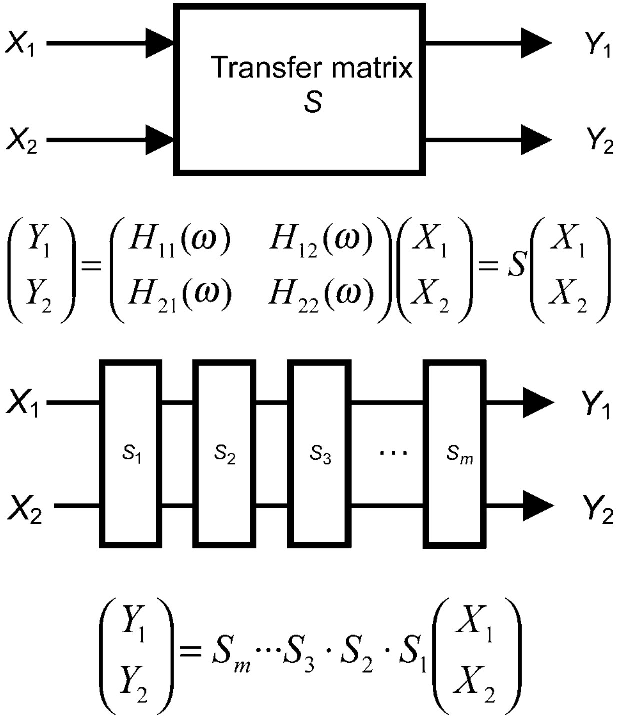
$$\mathrm{FSR} = \frac{\lambda^2}{nL}$$

B7.2.5 Optical modelling—beam propagation methods and the transfer matrix formalism

For the modelling of simple components or optical integrated circuit building blocks, the beam propagation method (BPM) and mode solvers are used:

- In BPM, the field is propagated stepwise through slices of the known waveguide structure, and in each step a phase correction is applied if the propagation constant of the guided mode changes through variations of the refractive index or the waveguide geometry. BPM is extensively treated in [20], [7] and [2].
- Mode solvers calculate the field distributions and propagation constants for the electrical and magnetical eigenmodes, and are often used as starting conditions for a subsequent beam propagation modelling. Methods solving the full Maxwell vector equations or semivectorial approximations are employed. The accuracy of different methods is compared in [7].

The applicability of the BPM is limited by the computation time needed for the simulation, so that alternative tools are needed for circuits with a higher degree of complexity. One way to achieve this is to use a transfer matrix formalism relating the amplitudes at the inputs and outputs of a device (figure B7.9), and the unitarity of the matrix assumes that filters are lossless. $H_{11}(\omega)$ is the complex transfer function giving access to the filter's power transmission $|H_{11}(\omega)|^2$ and phase $\phi(\omega) = \arctan[\mathrm{Im}(H_{11}(\omega))/\mathrm{Re}(H_{11}(\omega))]$.



$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} H_{11}(\omega) & H_{12}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = S \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = S_m \cdots S_3 \cdot S_2 \cdot S_1 \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Figure B7.9. Device with two input and output ports and unitary transfer matrix S (top); serial filter through concatenated two port devices (bottom).

Table B7.1. Transfer matrices and transfer functions for waveguide devices described in sections B7.2.3 and B7.24.

Sketch	Matrix/transfer function	Interpretation
Directional coupler	$\begin{pmatrix} \cos(\theta) & -\mathrm{j}\sin(\theta) \\ -\mathrm{j}\sin(\theta) & \cos(\theta) \end{pmatrix} \equiv \begin{pmatrix} \cos(\kappa z) & -\mathrm{j}\sin(\kappa z) \\ -\mathrm{j}\sin(\kappa z) & \cos(\kappa z) \end{pmatrix}$	$\theta =$ coupling angle ($\pi/4$ for 3 dB coupler)
Phase shifter/delay line	$\begin{pmatrix} \mathrm{e}^{-\mathrm{j}\phi} & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \mathrm{e}^{-\mathrm{j}2\pi nL/\lambda} & 0 \\ 0 & 1 \end{pmatrix}$	Propagation of a mode with effective index n, wavelength λ through a section of length L
Mach–Zehnder	$1/\sqrt{2}\begin{pmatrix} 1 & -\mathrm{j} \\ -\mathrm{j} & 1 \end{pmatrix} \times \begin{pmatrix} \mathrm{e}^{-\mathrm{j}\phi} & 0 \\ 0 & 1 \end{pmatrix} 1/\sqrt{2}\begin{pmatrix} 1 & -\mathrm{j} \\ -\mathrm{j} & 1 \end{pmatrix}$	For 3 dB couplers, $\theta = \pi/4$
Phasar/AWG	$H(\omega) = \sum_{l=1}^{N} A(l)\exp\left(-\mathrm{j}\omega \frac{n_{\mathrm{eff}} l \Delta L}{c}\right)$	$N =$ number of waveguides in the arrayed waveguide grating. $A(l)$ is the amplitude in guide l, following a Gaussian distribution
Open ring resonator	$H(\omega) = \frac{\rho - \exp[-\mathrm{j}(\omega T + \phi)]}{1 - \rho\exp[-\mathrm{j}(\omega T + \phi)]} = \frac{\cos(\theta) - \exp[-\mathrm{j}(\omega T + \phi)]}{1 - \cos(\theta)\exp[-\mathrm{j}(\omega T + \phi)]}$	ρ is the through path amplitude transmission of the coupler. T is the roundtrip time through the feedback loop
Parallel lattice filter/tapped delay line filter (MA)	$H(\omega) = \sum_{k=0}^{N-1} h_k \mathrm{e}^{-\mathrm{j}kT_k\omega}$	$N =$ number of delay lines. T_k is the time delay through arm k.

Transfer matrices and transfer functions of some devices described in this section are listed in table B7.1. The amplitude and phase characteristics of a more complex filter are then obtained from the multiplication of the elementary matrices, and examples are the serial and parallel lattice filters to be described in section B7.5.

Finally, the transfer function can be expressed as

$$H_{11}(\omega) = \frac{A(\omega)}{B(\omega)}$$

where the polynomials $A(\omega)$ and $B(\omega)$ have only zeros. Depending on the numbers of zeros or poles of $H_{11}(\omega)$, filters are classified as [21, 22]:

- Moving average (MA), also called finite impulse response (FIR), filters have only zeros and consist only of forward paths. Examples are the AWG and the Mach–Zehnder interferometer.
- Autoregressive filters (AR) have only poles, and the two-port ring resonator (figure B7.8(*a*)) is one example.
- Autoregressive moving average filters (ARMA) have both zeros and poles. Infinite impulse response filters (IIR) represent a sub-category with at least one pole.

For a detailed treatment of the transfer matrix formalism, see references [20] (chapter 6.5) and [22] (chapter 3.3), with application to the synthesis of optical filters in [23], [24] and [25].

B7.3 Integrated optics materials and fabrication technology

A major challenge facing integrated optical circuit developers is the fact that devices can be fabricated in different materials, and the ultimate technology choice will be based on the performance, manufacturability and cost.

Hybrid integration (section B7.4) allows combining the best suitable materials on one platform, but for monolithic circuits a compromise may be needed.

This section will give an overview of the commonest materials, their properties and the fabrication processes currently deployed in integrated optics; we refer to chapter A1.2 for further reading.

B7.3.1 Silica

Most of the commercially available passive opto-electronic circuits use doped silica waveguides, as propagation losses below $0.1\,\mathrm{dB\,cm^{-1}}$ are easily achieved in the 1.3 and $1.5\,\mu\mathrm{m}$ telecom windows. Pioneering work in this field has been done in the early 1970s–1980s by NTT, and today devices are commercialized by many vendors (NEL, SDL/PIRI, IONAS/NKT Integration, Agere, Hitachi etc).

A typical fabrication process based on deposited thin films is shown in figure B7.10: a thin layer of silica is deposited on a planar silica or silicon substrate via chemical vapour deposition (CVD) processes such as flame hydrolysis (FHD) [26, 3] or plasma enhanced deposition (PECVD) [27] of gaseous precursors. In the case of the silicon substrate, a pure SiO_2 buffer layer with $10–20\,\mu\mathrm{m}$ thickness optically isolates the waveguide layer. The index of the waveguide layer is raised from the initial silica index ($n = 1.444$ at 1550 nm) above the index of the cladding by adding dopants such as germanium, phosphorus or titanium during thin film deposition, and index contrasts between $0.34\%\Delta$ (matched to standard single mode fibre) and $2\%\Delta$ are obtained with losses as low as $0.017\,\mathrm{dB\,cm^{-1}}$ for $0.75\%\Delta$ [26]. Mechanical stresses built up during waveguide deposition can be released by annealing at high temperature [28], and stress due to the mismatch of the thermal expansion coefficients of the substrate and waveguide materials can be avoided by a proper choice of dopant levels and waveguide geometries [29, 30]. Waveguide patterning is done by photolithography and etching [31, 32 chapter V-2], and finally a cladding covers the waveguides. Additional metallization layers can be deposited for the powering of thermo-optic phase shifters ($\mathrm{d}n/\mathrm{d}T = 1\times10^{-5}\,\mathrm{K^{-1}}$ [3]).

An alternative to thin film deposition is to fabricate the waveguides through ion exchange [33]: alkali ions, such as Na^+ or K^+, present in the glass are replaced by another cation such as Ag^+ or Tl^+, which locally increase the index proportional to their concentration. The process for fabricating buried channel waveguides usually consists of two steps: first, a glass substrate with an appropriate lithographic mask is immersed in a molten salt containing the silver or thallium ions, but no sodium or potassium ions. There will be an inter-diffusion of the species, resulting in a channel waveguide at the surface of the substrate. In the second step, an electric field is applied leading to the migration of the exchanged ions

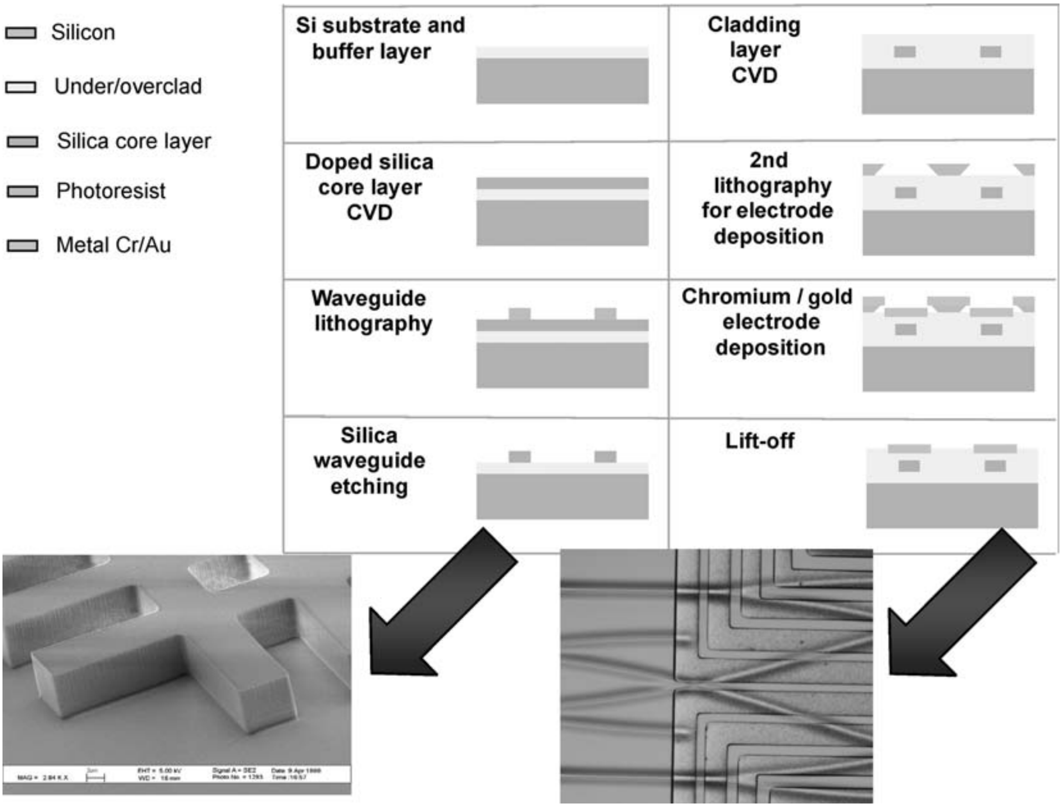

Silicon
Under/overclad
Silica core layer
Photoresist
Metal Cr/Au
Si substrate and buffer layer
Doped silica core layer CVD
Waveguide lithography
Silica waveguide etching
Cladding layer CVD
2nd lithography for electrode deposition
Chromium / gold electrode deposition
Lift-off

Figure B7.10. Fabrication process for silica waveguides [198]. The buffer layer separates the core layer from the substrate. Waveguides are formed via etching, covered by a silica cladding layer, and electrodes for tunable components are usually deposited via physical vapour deposition.

deeper into the substrate. The resulting waveguides have an index profile, which is determined by the diffusion characteristics of the Ag^+ or Tl^+ ions. Passive directional couplers with $<0.15\,\mathrm{dB\,cm^{-1}}$ propagation loss have been reported in [34].

At present, ion exchange is primarily used for fabricating waveguides in lithium niobate (B7.3.4) and rare-earth doped glasses (section B7.3.6).

The availability of strong laser sources has allowed us to study the photo-refractivity in glasses: in GeO_2 doped films, permanent index changes can be induced by irradiating with 240 nm [35] or 157 nm laser light [36], and the sensitivity is usually enhanced by deuterium loading [37]. Index changes Δn between -6×10^{-3} [38, 39] and $+3\times10^{-3}$ [37] have been reported, allowing direct laser writing of waveguides [35], Bragg gratings [40] and waveguide couplers [41].

B7.3.2 Silicon oxynitride

Higher index contrasts and thus integration densities are available from silicon oxynitride (SiO_xN_y or SiON) films, which cover the index range between SiO_2 ($n = 1.45$) and Si_3N_4 ($n = 2.0$) and are deposited by PECVD or LPCVD (low pressure CVD) [27]. Although not yet commercialized in volume, there is an industry pull mainly from IBM and Kymata/Alcatel.

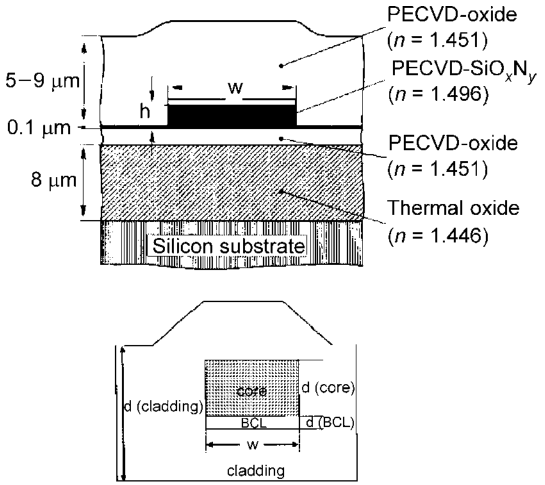


Figure B7.11. PECVD waveguides with compensated birefringence: rectangular waveguide cross section [42] (top, indices are at 1300 nm); guide with Si_3N_4 birefringence compensating layer BCL underneath the channel waveguide [43] (bottom). Reproduced by permission of IEEE.

Reduction of the stress-related birefringence after waveguide fabrication has been reported by having waveguides with a rectangular cross-section [42], or by an extra Si_3N_4 birefringence compensating layer underneath the channel waveguide [43] as shown in figure B7.11.

Propagation losses of $<0.1\,\mathrm{dB\,cm^{-1}}$ are achievable for lower index oxy-nitride materials, and thermo-optic devices (the coefficient $\mathrm{d}n/\mathrm{d}T = 1.2\times10^{-5}\,\mathrm{K^{-1}}$ being close to that of silica) have been made out of SiON waveguides with $3.3\%\Delta$ index contrast [44]. In this particular case, coupling losses due to the mismatch of the mode field diameter have been reduced by attaching a fibre with smaller core diameter.

Silicon-rich nitride (SRN) [45] deposited by LPCVD allows us to obtain index contrasts of $\Delta n = 0.6$, and thus a bending radii of 40 μm. Although the reported propagation losses are still of the order of $0.6\,\mathrm{dB\,cm^{-1}}$ (measured in a one-dimensional slab waveguide), SRN will be a promising candidate for high density integrated optics.

B7.3.3 Silicon-on-insulator

Silicon-on-insulator (SOI) integrated photonic circuits may have a potential for true monolithic integration with electronics due to their compatibility with the micro-electronics CMOS process, which can provide gigabits/second electronic circuitry with a low noise. As we will see in section B7.4, silicon is also the preferred substrate material for hybrid integrated circuits. The main commercial source for SOI devices is Bookham (transceivers, multi-channel monitors).

SOI layers can be fabricated by the bonding of thermal oxide layers (BESOI [46]), oxygen ion implantation [47, 48] and sputtering, CVD or evaporation [49].

Single mode rib waveguides [50] with dimensions comparable to single mode fibre are fabricated via etching techniques, and propagation losses of $<0.2\,\mathrm{dB\,cm^{-1}}$ [51] at 1550 nm and $<0.1\,\mathrm{dB\,cm^{-1}}$ at 1300 nm [52] are obtained. The cross-section of a SOI rib waveguide is shown in figure B7.12.

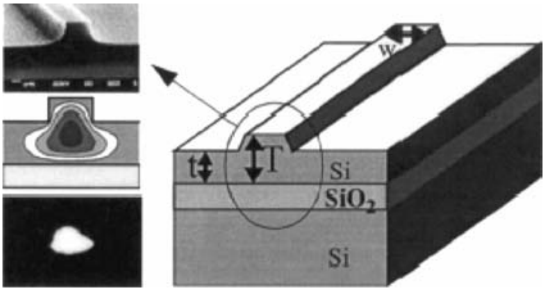


Figure B7.12. Silicon-on-insulator (SOI) single mode rib waveguide [46] fabricated by wafer bonding and etching. Schematic (right), scanning electron micrograph (left top), modelled (left middle) and experimental guided mode (left bottom). Reproduced by permission of IEEE.

The high index of $n = 3.45$ allows high confinement of the optical mode. This makes Si a suitable material for photonic bandgap structures [53], for which an index contrast of >2 is required for guiding [54].

Modulators have used the thermo-optic coefficient ($\mathrm{d}n/\mathrm{d}T = 1.8 \times 10^{-4}\,\mathrm{K}^{-1}$ [5]). Carrier injection is the most important electro-optic effect in Si [55, 56], and index changes of the order of $\Delta n = 1.5 \times 10^{-3}$ have been demonstrated [57, 48].

B7.3.4 Lithium niobate

Lithium niobate ($\mathrm{LiNbO_3}$, often abbreviated as LN) is a uniaxial crystal with large birefringence ($n_{\mathrm{extraordinary}} = 2.2$, $n_{\mathrm{ordinary}} = 2.286$) and a large linear electro-optic (Pockels) effect of $r_{33} = 30.9\,\mathrm{pm\,V^{-1}}$ [9]. Wafers with 100 mm diameter are commercially available [58]; common techniques for fabricating waveguides are thermal in-diffusion of titanium or proton exchange [59] and both processes do not seriously disrupt the lattice structure of the host material.

Ti-diffused channel waveguides typically have index contrasts of $\Delta n = 0.5$–1% [60] and $<0.1\,\mathrm{dB\,cm^{-1}}$ propagation losses [61], but index contrasts of up to 7% are possible with proton exchange [59].

The large EO effect of LN is used for external modulation of laser sources, and bandwidths of up to 100 GHz are reported on probed devices [62]. For achieving these high modulation bandwidths, special designs for the electrodes are required to match the velocities of the electrical and optical mode [58]. Modulators are treated in detail in chapter B4. Main industry players in the modulator segment are JDS Uniphase and Corning OTI.

Polarization converters and tunable wavelength filters via acousto-optic effects are reported in [63 chapter B7.8].

B7.3.5 Polymer

While not traditionally a high performance optical material, polymer has a number of attributes that make it an interesting material for optical circuits [64] and structural devices [65].

For optical waveguides, photosensitive optical polymers based on various monomers (including acrylates, polyimides, cyclobutenes) have been developed and are commercially available. Upon exposure, these monomers form crosslinked networks and materials with indices between 1.3 and 1.6 are available [4]. Mixing of two polymers which can be co-polymerized allows tailoring of the refractive index with an accuracy of 10^{-4}.

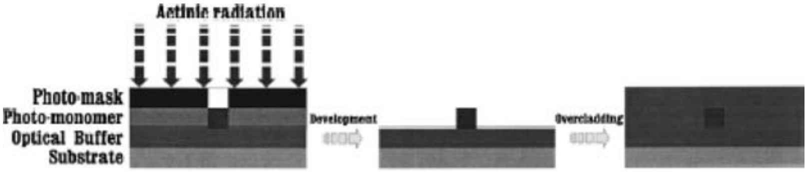


Figure B7.13. Schematic of the photolithography process for patterning of polymer waveguides [64]. Reproduced by permission of SPIE.

Adding photoinitiators to the material allows fabrication of waveguides by conventional mask photolithography (figure B7.13) or direct waveguide writing, although etching or moulding processes are also possible [4].

Standard materials offer waveguide propagation losses of 0.2 dB cm^{-1}, and the large temperature dependence of the refractive index $\mathrm{d}n/\mathrm{d}T = -2 \cdots - 3 \times 10^{-4}\,\mathrm{K}^{-1}$ [4] makes polymers particularly attractive for thermally tunable devices with low power consumption such as photoinscribed Bragg gratings and Mach–Zehnder switches, but also passive AWGs [66]. Main commercial products are thermo-optic switches (Akzo Nobel/JDS Uniphase).

The wide index range allows us to use polymers in hybrid silica/polymer devices [67], and the negative $\mathrm{d}n/\mathrm{d}T$ can compensate the positive $\mathrm{d}n/\mathrm{d}T$ of silica for athermalizing temperature sensitive silica waveguide components [68, 69, 70].

Electro-optic coefficients of $r_{33} = 10–15\,\mathrm{pm\,V^{-1}}$ have been achieved in polymer waveguides containing chromophores aligned in an electric field. Although photochemical stability and loss (typically 1 dB cm^{-1}) still remain to be addressed, modulating devices with >110 GHz bandwidth have been reported [71].

Waveguides with integrated structures to grip optical fibres with $< 1\,\mu$m alignment tolerance can be fabricated by etching [64], but also LIGA moulding processes [65], and optical backplanes have been reported in [72].

B7.3.6 Active waveguides

Active waveguides have been obtained from the implantation of rare-earth ions into a host material [73], ion exchange in Er/Yb co-doped glasses [74] and through PECVD thin film deposition [75].

Er^{3+} has a transition in the 1.54 μm telecom window, and optical pumping from the ground to the excited states is possible by 980 or 1480 nm pump lasers (figure B7.14) [73]. Net gains of 20 dB can be obtained with 120 mW pump power [63 chapter 6] (see this reference for a state-of-the-art on this subject). TEEM Photonics and NKT Integration commercialize Er^{3+}-doped planar amplifiers.

In a similar way, Nd^{3+} doping in silica waveguides provides a gain at 1.05 nm wavelength [76].

Active Er-doped devices have also been fabricated in $\mathrm{LiNbO_3}$ [77, 7 chapter 6], and other oxide materials, ceramics and Si [78].

B7.3.7 Indium phosphide gallium arsenide

Among the III-V semiconductors, InP plays a major role, as it is the current material of choice for the fabrication of lasers and detectors, which in turn can be monolithically integrated with passive InP waveguides. The possibility of matching the lattice constant of the quarternary alloy $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}_{1-y}\mathrm{P}_y$ to InP over a large window of compositions results in a wide range of bandgaps, thus operating wavelengths between 1000–1700 nm are obtained. However, the fabrication process is quite expensive and wafer sizes are limited to 2–3 inches. Although we will illustrate a few monolithic InP circuits

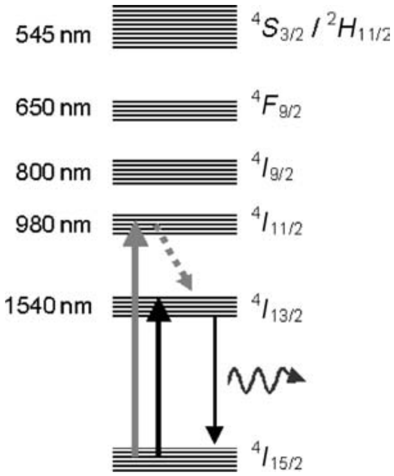


Figure B7.14. Simplified representation of the Er^{3+} energy levels. Optical pumping into the excited $^4I_{11/2}$ and $^4I_{13/2}$ states requires semiconductor lasers operating at 980 or 1480 nm [73]. Reproduced by permission of the Materials Research Society.

integrating passive and active device functionalities in section B7.5, most of the InP-based components today are combined with other passive components through hybrid integration (section B7.4.).

GaAs has a major role in the development of high speed electronics, thus up to 6 inch wafers are available. The bandgap of GaAlAs/GaAs and InGaAs/GaAs can be engineered for obtaining laser operation in the 780 and 980 nm wavelength regions, respectively, the latter being used for optical pumping of Er^{3+}-doped glasses.

We refer to chapters B1.1, B1.2 and B2 for further reading.

B7.3.8 Conclusion

Table B7.2 summarizes the key properties and primary phase tuning mechanism of the materials presented in this section. Each of the materials has its advantages and disadvantages, so that there is no clear winner. The choice of the material depends primarily on the specific needs for a given functionality.

B7.4 Device packaging and function integration

The first part of this section is about packaging of opto-electronic circuits, which fulfil a single or multiple functions in the network. The second part will discuss briefly the different technologies for the integration of multiple functions on one chip, and we will focus more specifically on hybrid integration, as this is the main integration technique today.

B7.4.1 Device packaging

There are different levels of opto-electronic integration in a communication system [79] (figure B7.15):

- A system consists of several racks or cabinets of equipment.
- A rack or cabinet contains frames, which in turn have units combining a multitude of optical and electronic functions.
- A single or a few optical functions together with control electronics are packaged on a board.

Table B7.2. Key properties of the materials described in section B7.3. For the coupling loss to a standard single mode fibre SMF28 [82] we indicate only a range (high >3 dB/interface, medium 1–3 dB/interface, low <1 dB/interface), as exact coupling loss values (see section B7.4.1 and table B7. 3) depend on mode field diameters which in turn can be adapted.

	Propagation loss	Coupling loss	Refractive index n	Primary phase tuning mechanism	Primary application
Silica	<0.05 dB cm^{-1}	Low–medium	1.444	Thermo-optic	Passive components, thermo-optic tunable devices
Silicon oxynitride	<0.1–0.2 dB cm^{-1} for n around 1.45 0.6 dB cm^{-1} for n around 2.0	Low–medium	1.444–2.0	Thermo-optic	Passive components, thermo-optic tunable devices
Silicon-on-insulator SOI	<0.2 dB cm at 1550 nm	High	3.45	Thermo-optic, electro-optic	Passive components, photonic bandgap devices
Lithium niobate LN	<0.1 dB cm^{-1}	Medium	2.0	Primarily electro-optic, but also thermo-optic and acousto-optic	High speed gigabit/second modulators and switches
Polymer	<0.2 dB cm^{-1}	Low–medium	1.3–1.6	Thermo-optic, electro-optic (chromophores)	Passive components, Thermo-optic tunable devices/switches with low power consumption
Er^{3+} doped glasses	<0.1 dB cm^{-1}	Low	1.49–1.50		Pre- and post-amplifiers
InP	ca. 3 dB cm^{-1}	High	3.1	Electro-optic, $\chi(3)$ nonlinearity	Lasers, detectors, semiconductor optical amplifiers, gigabit/second optical gates and switches
GaAs	ca. 0.5 dB cm^{-1}	Medium	3.3737		Pump lasers (980 nm)

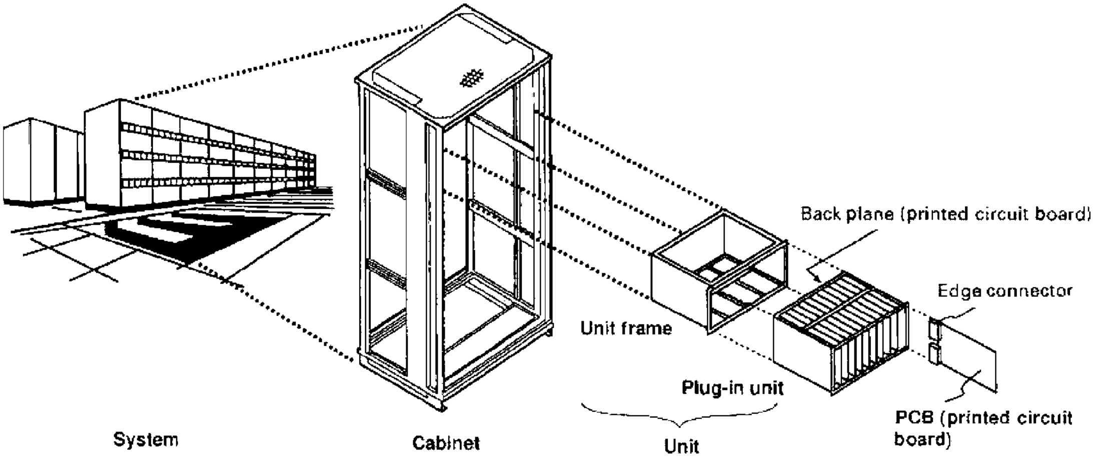

Back plane (printed circuit board)
Edge connector
Unit frame
Plug-in unit
PCB (printed circuit board)
System
Cabinet
Unit

Figure B7.15. Packaging levels in a generic communication system [79]. Reproduced by permission of John Wiley & Sons, Inc.

The purpose of packaging is (as an example, see figure B7.16)

1. (1) To connect a fibre to the monolithic or hybrid integrated optic *chip* without adding loss.
2. (2) The monolithic or hybrid integrated optical circuit can in turn contain dynamic elements (phase shifters, active opto-electronic devices) which need to have an *electrical interconnection to control electronics*.
3. (3) The fibre-optical circuit ensemble needs to be fixed into a *housing* or *package* without degradation of the mechanical stability or optical performance.
4. (4) The housing or package ensures *reliability* under different operating or environmental conditions.

Coupling from a fibre to an optical integrated circuit can be accomplished [80]

- From the waveguide surface via prisms or surface gratings.
- Through edge coupling into the waveguide cross-section via focusing lenses, or through direct attachment of the fibre to the chip. For the latter, fibres can be placed into V-grooves fabricated into Si substrates [81], or actively aligned to the chip which is the standard method used today.

Edge coupling losses arise from the mismatch of the fibre and waveguide mode fields, as well as from angular or lateral misalignments and are given by the overlap integral

$$\text{Loss} = -10\,\text{dB} \times \log_{10}\left[\frac{\left(\int_{\text{area}} \phi_1 \phi_2 \mathrm{d}A\right)^2}{\left(\int_{\text{area}} \phi_1^2 \mathrm{d}A\right)\left(\int_{\text{area}} \phi_2^2 \mathrm{d}A\right)}\right]$$

with ϕ_1, ϕ_2 being the mode fields and A the waveguide cross-section. Approximations for the various contributions are given in table B7.3.

As an example, the mode-field mismatch of a standard single mode fibre (mean field radius $w_0 = 4.6\,\mu\text{m}$ [82]) to a planar waveguide ($w_0 = 3.3\,\mu\text{m}$) results in losses of the order of 0.5 dB. To keep

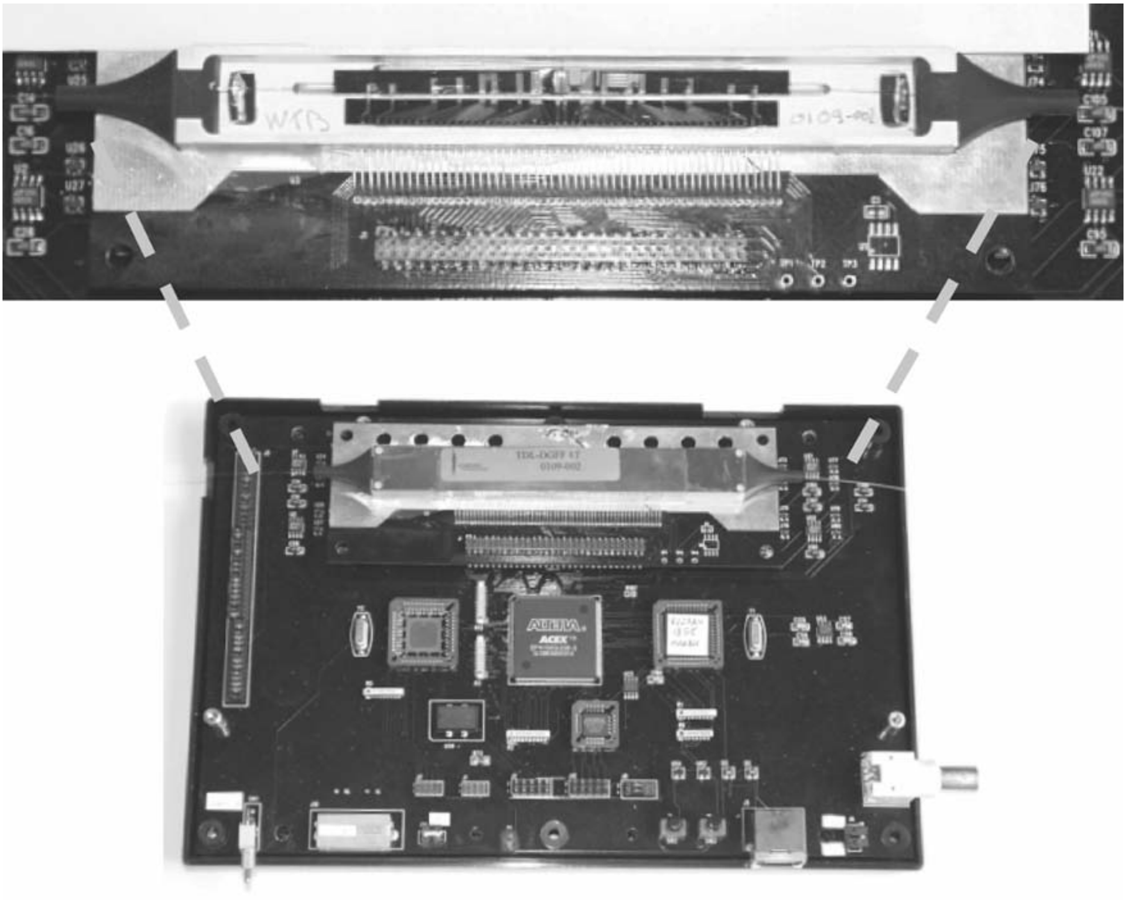


Figure B7.16. Example design for a opto-electronic component consisting of a silica waveguide chip in its housing (top) and electronics control board [135]. Reproduced by permission of IEEE.

the excess loss from lateral misalignment to below 0.3 dB, the fibre needs to be positioned with $<1.0\,\mu\text{m}$ accuracy. The fibres are usually attached by optical adhesives or directly fused to the chip using a CO_2 laser [83, 84].

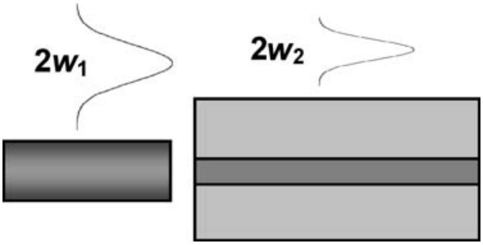
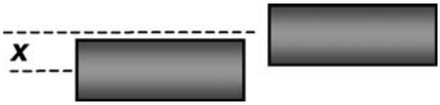
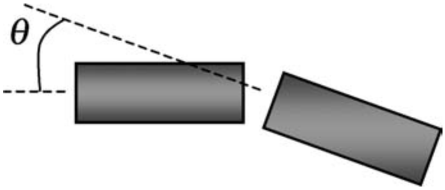
Edge coupling of III–V components such as lasers, detectors or semiconductor optical amplifiers (SOAs) to fibres is more difficult, because their small modefield diameter of 1–2 μm would immediately lead to increased mode-field mismatch losses and also slightly reduced alignment tolerances.

In this case, loss reduction is possible by using fibres with anamorphic lenses [85] or enlarging the mode field of the active component with a taper [86] (figure B7.17) [87]. These relaxed alignment tolerances are the keys to high yield low cost passive alignment techniques as used in hybrid integration.

For electrical interconnection of active semiconductor and dynamic elements mainly two technologies are used in opto-electronics [88, 89, 90]:

- Bonding of a thin (typically 10–250 μm diameter) gold or aluminium wire onto metallic pads near the OE circuit chip edge. Wire bonding is a sequential process, and the wire length of a few millimetres leads to parasitic impedances limiting the electrical bandwidth to about 15 GHz.

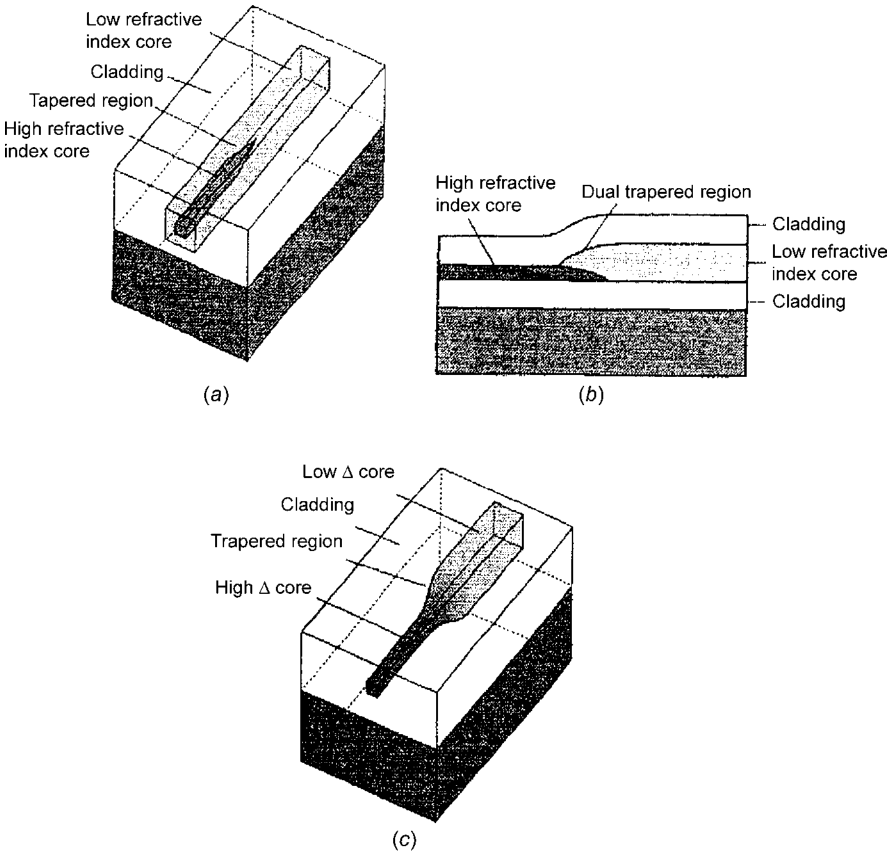
Table B7.3. Coupling loss mechanisms and formulas.

Loss mechanism	Loss (dB)	Schematic
Mode mismatch	$-20\log_{10}\left[\frac{2w_1\cdot w_2}{w_1^2+w_2^2}\right]$ w = mean field radius (1/e amplitude)	
Transverse offset	$-10\log_{10}\left[\frac{x^2}{w^2}\right]$ $\approx 4.343\left(\frac{x}{w}\right)^2$ where $w=\sqrt{(w_1^2+w_2^2)/2}$ and x is the offset	
Angular misalignment	$-10\log_{10}\left[\exp\left[-\frac{1}{2}\left(\frac{2\pi n_{\mathrm{eff}}w\theta}{\lambda}\right)^2\right]\right]$ $\approx 8.69\left(\frac{\pi n_{\mathrm{eff}}w\theta}{\lambda}\right)^2$ where $w=\sqrt{(w_1^2+w_2^2)/2}$ and θ is the angular misalignment (rad)	

- Shorter interconnection (typically 100 μm), and thus higher electrical bandwidth, is possible with the flip chip process, which is moreover a parallel process. The electrical interconnection can be made with higher density anywhere on the device surface, and the flip chip (or solder bump self-alignment technique) is also one of the passive assembly techniques used in hybrid integration.

The package or housing of the device needs to enclose and protect the optical circuit without performance degradation through

- Thermal effects—heat is generated by the active components, so that a good dissipation is essential for maintaining the circuit at stable temperature. This is achieved by


Low refractive index core
Cladding
Tapered region
High refractive index core
(a)
High refractive index core
Dual trapered region
Cladding
Low refractive index core
Cladding
(b)
Low Δ core
Cladding
Trapered region
High Δ core
(c)

(a) Dual core waveguide
(b) Dual trapered core waveguide
(c) Trapered core waveguide

Figure B7.17. Mode-field converting waveguides [86]. The principle is to have the high index waveguide (with small mode-field diameter) connected through an intermediate taper to a low index waveguide (with larger mode-field diameter), which is then coupled to a single mode fibre. Reproduced by permission of John Wiley & Sons, Inc.

  - * Having laser submounts with high conductivities (diamond, silicon carbide (SiC), alumina (Al_2O_3), aluminium nitrite (AlN), silicon (Si)) or Si substrates for thermo-optic and hybrid components.
  - * Thermo-electric coolers or passive heat dissipation.
- Stress related effects, as they induce birefringence and change the response of an optical circuit based on phase control. Potential issues are addressed by using substrates and packaging materials with adapted thermal expansion coefficients and filled adhesives with low Young's modulus.

At present, commonly used housing materials are aluminium, Kovar, Invar and organic/inorganic composites with near zero expansion coefficient [79]. Bending losses or breakage of the fibre attached to the optical circuit can be avoided by keeping the fibre bend radii above 2–3 cm.

Opto-electronic components are required to have failure-free performance over a range of operating temperatures (typically −5 to +70°C) over their lifetime (miminum 25 years). For assuring a reliable operation, testing procedures are defined [91], the purpose of which is to

- Verify reliability against thermal shocks, thermal cycles (−40 to 70°C), damp heat (85°C/85% relative humidity) and moisture.
- Screen devices through thermal cycling and short term storage at high temperature.
- Estimate life expectancy through accelerated ageing (70 or 85°C, 2000–5000 h). The principle is to test several identical components at different temperatures and to estimate the mean time to failure (MTTF) from an Arrhenius relation.

Depending on the application, operating environment and chip materials, the OE components can be packaged hermetically or nonhermetically.

B7.4.2 Function integration and hybrid technology

The integration of optical functions can be accomplished mainly in three different ways:

(1) Precision placement of the various elements into one package and interconnection via free-space optical links. This method has limitations for the complexity of the OE circuit, and the cost reduction potential through mass production is questionable.

(2) In monolithic integration, the optical and electrical functions are fabricated—in analogy to electronic LSI—on the same semiconductor substrate such as InP or GaAs. The main drawback is that currently best performances for different optical functions may be obtained with different materials (see section B7.3), for example modulators with $LiNbO_3$, passive wavelength multiplexers with silica waveguides and so on. Also, the achievable integration density is limited by optical constraints (waveguide bend radii and the size of the fibre/chip interconnection) rather than the size of the electronic circuitry.

(3) In hybrid integration, active devices and electronics are assembled together with passive lightwave optical circuits on a common platform, which allows combining the best available circuits. Furthermore, the yield for this technology may be higher than for monolithic integration, since the components can be selected for performance prior to the assembly. The remainder of the section reviews the critical technologies needed for surface-hybrid integration (for the various types of hybrid integration, see reference [86]).

A generic hybrid integrated circuit is shown in figure B7.18, and the main technical ingredients for functional integration are:

(1) A platform onto which the active and passive optical circuits are integrated. Among the materials used in packaging (diamond, SiC, Al_2O_3, AlN, Si), Si is considered to be the best material for a number of reasons: low cost, good thermal conductivity ($1.57\,\mathrm{W\,cm^{-1}\,K^{-1}}$) for evacuation of thermal load from active components, a thermal expansion coefficient ($\mathrm{CTE} = 2.33 \times 10^{-6}\,\mathrm{K^{-1}}$) almost matched to that of InP components ($\mathrm{CTE} = 4.5 \times 10^{-6}\,\mathrm{K^{-1}}$), availability of processes for fabrication of electrical circuits and Si alignment features. Moreover, passive silica waveguide components are usually fabricated on Si substrates.

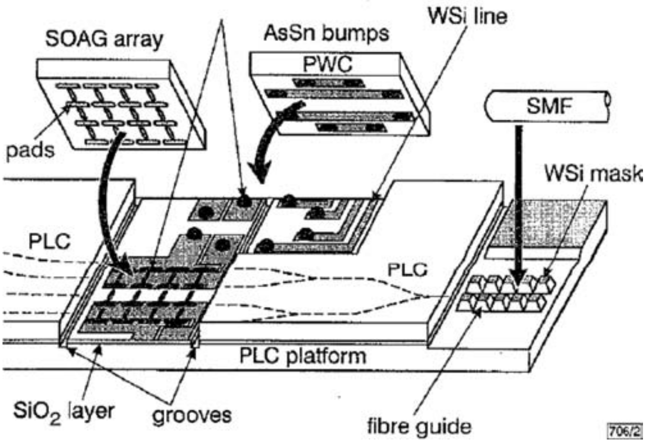

SOAG array
AsSn bumps
WSi line
PWC
SMF
pads
WSi mask
PLC
PLC
PLC platform
SiO$_2$ layer
grooves
fibre guide
706/2

Figure B7.18. Hybrid integrated circuit [168] showing passive planar circuits (PLC), electronic transmission WSi lines, single mode fibre (SMF) with passive fibre guides and active circuits (semiconductor optical amplifier gates SOAG) with AuSn solder bumps for flip chip self-alignment. Reproduced by permission of IEEE.

There are mainly two different types of platform:

(i) The Si optical bench has fibre V-grooves fabricated by anisotropic etching as well as other alignment features for active components, but has no passive waveguides.

(ii) NTT has developed a planar lightwave circuit (PLC) platform [92, 93], which also integrates passive silica waveguide components. A cross-section of the structure is depicted in figure B7.19.

(2) In the integration step, the optical building blocks need to be interconnected with low optical loss due to physical misalignment, i.e. within the roughly 1 μm accuracies mentioned in section B7.4.1. Both active and passive alignment methods [63 chapter 10] are used today:

In an active alignment process—for example coupling of a fibre to a laser diode—the optimum position is found through a measurement of the transmitted laser power, and sub-micrometre accuracy can be achieved with commercial precision alignment stages.

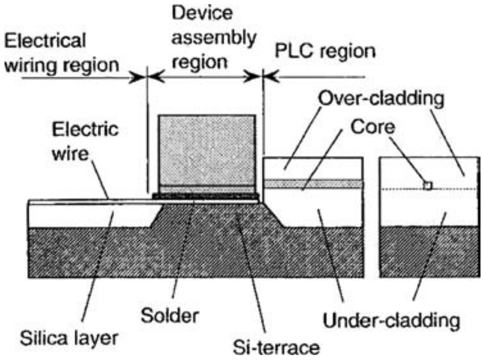

Electrical wiring region
Device assembly region
PLC region
Over-cladding
Core
Electric wire
Silica layer
Solder
Si-terrace
Under-cladding

Figure B7.19. Cross-section of the NTT PLC platform: the Si terrace and the deposited silica buffer layer are planarized, thus defining the vertical plane for subsequent passive assembly of the active components [63]. Reproduced by permission of Springer-Verlag.

In a passive alignment process, the accuracy is defined through alignment features fabricated on both the OE building blocks and the integration platform:

(i) The index method uses microscopes and precision manipulators for accurate alignment in the horizontal (platform) plane; an example is shown in [94].

(ii) Mechanical contact alignment uses topographic features such as stops, standoffs, notches, pedestals (figure B7.20) for alignment in both horizontal and vertical directions. The major difficulty is the precise definition of the vertical plane during the waveguide fabrication process. Accuracies of 0.8 μm are reported for lasers coupled to fibres [95], passive alignment of single mode fibres to both facets of a semiconductor optical amplifier is shown in [96], and integration of both diode laser and photodiode with <3 μm precision has been demonstrated in [97].

(iii) The flip chip or solder bump technique uses precisely known volumes of solder confined between wettable metal pads on either side of the bond. Raising the temperature above the solder melting point will simultaneously align the structures through surface tension and provide an electrical interconnection [98, 90] (figure B7.21). Solder bumps can be fabricated by vapour or liquid phase deposition [99], and a variety of active/passive circuits have been published with alignment accuracies of up to ± 1 μm [98].

(iv) A combination of mechanical and solder bump alignment has been reported for transceiver modules [100, 101].

(3) A reliable mechanical and electrical interconnection between the opto-electronic circuits and the platform. Materials and processes commonly used are [102, 103, 98]:

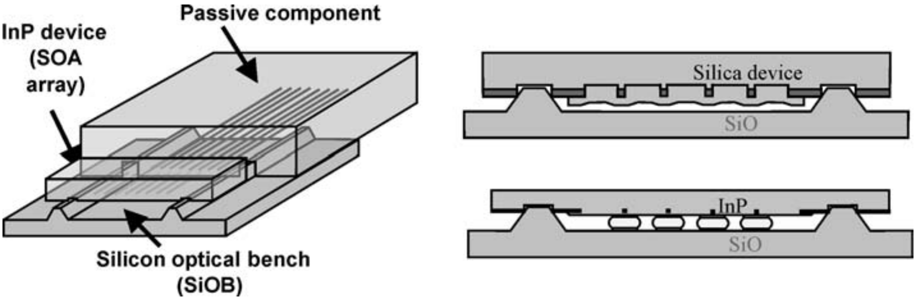


Figure B7.20. Mechanical contact alignment: notches fabricated into the active InP and passive silica waveguide components allow sub-micron alignment on a common silicon optical bench [199].

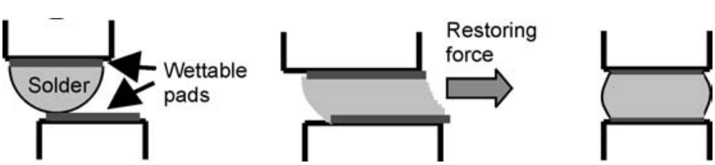


Figure B7.21. Flip chip solder bond—solder bumps with precise volumes (typically AuSn or PbSn) are fabricated on wettable pads, then brought into contact and heated above the melting point. The surface tension acts as a restoring force, resulting in alignment.

(i) In the flip chip process, eutectic 63%Sn/37%Pb (melting point 183°C) or 80%Au/20%Sn (melting point 281°C) solder bumps are fabricated on Ti or Cr adhesion layers.

(ii) Bonds can also be formed through a thermo-compression process at lower temperatures using Au/Sn or Au/In alloys.

(iii) Metal-filled epoxies with good thermal and electrical conductivity are suitable for curing at 150–200°C.

B7.5 Optical networks and integrated optical functions

With the increasing demand for communications bandwidth, digital optical transmission has become more and more important over the last few years. Currently, systems with capacities of hundreds of gigabits/second are commercially available, and transmission of 10.2 Tbit s^{-1} over 100 km of single fibre has been demonstrated in laboratory experiments [104]. In practice, this is done by simultaneously carrying optical signals with slightly different wavelengths over a single fibre (wavelength division multiplexing, WDM). In parallel, the bit rates for single channels have steadily increased to 40 Gbit s^{-1}, and in future systems of higher capacities will be achieved through time domain multiplexing (TDM) of several signals with identical wavelength into one high speed data stream.

The ultimate system figure of merit in digital transmission is the bit error rate (BER), which is defined as the probability of detecting a ‘1’ although a ‘0’ was received and vice versa. Typical goals are $\mathrm{BER} < 10^{-9}$ for voice traffic and $\mathrm{BER} < 10^{-12}$ for data transmission. For a fixed bit rate, the BER mainly depends on the received power and one speaks about power penalty (PP) when imperfect components are introduced into a network resulting in a BER degradation.

For optical networks, we refer to section C1, but for simplicity and for illustrating the use of integrated optical devices we will consider a generic WDM system as shown in figure B7.22:

- Light from a semiconductor *transmitter laser Tx* is *modulated*. In a WDM system, several signals are combined into the transmission fibre by an *optical multiplexer* MUX.
- The propagating signals are amplified every 70–100 km; mainly *Er^{3+} doped fibre amplifiers* are deployed [105], and distributed *Raman amplification* [106, chapter 8] in the transmission fibre is used to increase the system reach. Er^{3+} doped amplifiers need semiconductor pump lasers operating at 980 and 1480 nm; Raman amplified systems are pumped at 14xx nm wavelength.
- The gain of an EDFA is wavelength dependent, which leads to significant variations in signal-to-noise ratio (OSNR) of the received channels. The OSNR in turn depends on the power per channel and the number of amplifiers in the system. Nowadays, the channel power is periodically flattened with (static) thin film filters, but upcoming reconfigurable systems will require *dynamic gain equalization*.
- The finite spectral width of the source laser together with the positive fibre chromatic dispersion (i.e. the variation of the group velocity with wavelength) and nonlinear effects (self-phase modulation (SPM), cross-phase modulation (XPM), four-wave mixing) in the transmission fibre lead to a spreading of the optical pulse, which can be reshaped in *chromatic dispersion compensating devices*.

  In addition to that, the transmission fibre may not have perfect circular symmetry, giving rise to different propagation velocities as a function of the incoming polarization state (*polarization mode dispersion (PMD)*).
- More complex optical networks contain *add–drop multiplexers* (ADMs) and/or *optical crossconnect* (OXC) switches for interconnecting fibre links and for provisioning a path through the network:

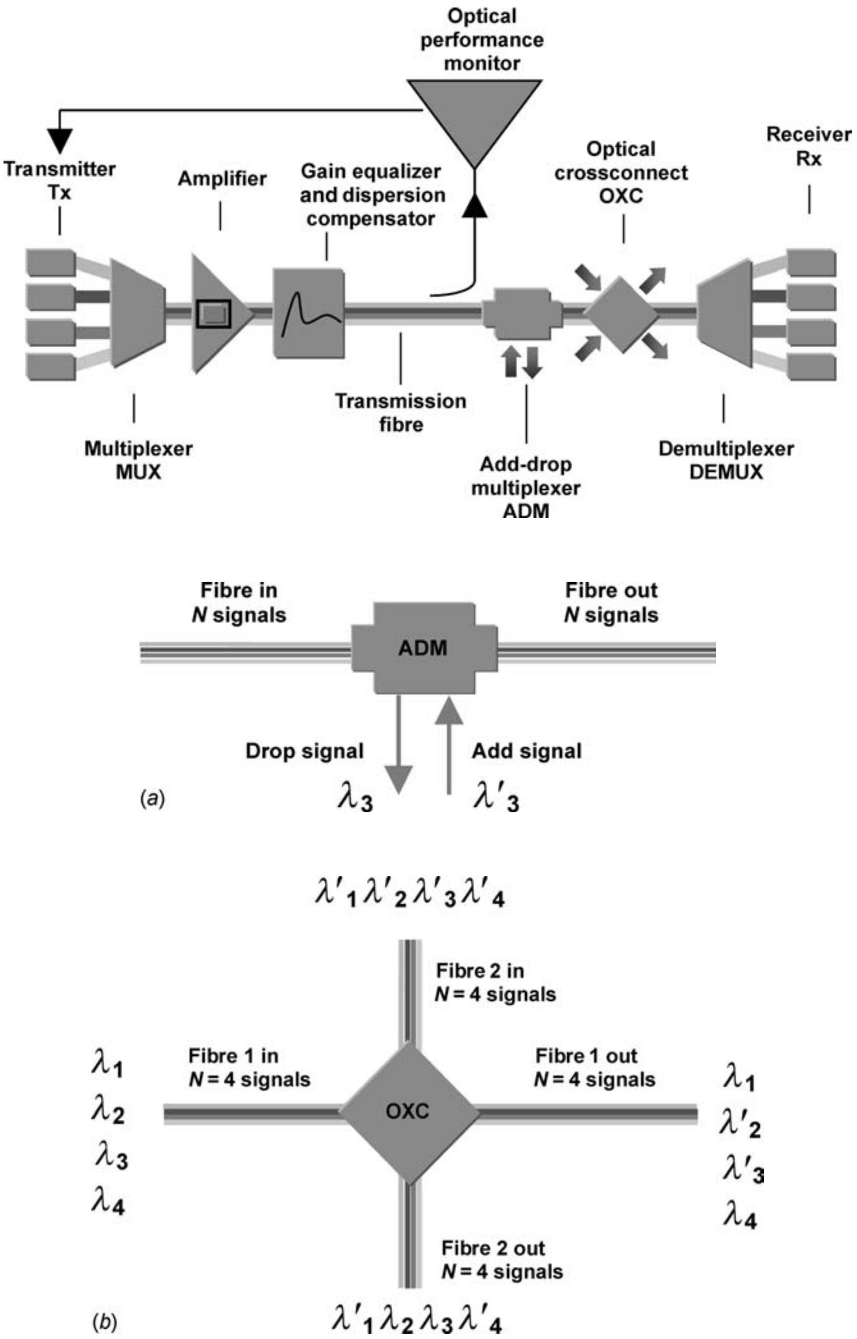

Optical performance monitor
Transmitter Tx
Amplifier
Gain equalizer and dispersion compensator
Optical crossconnect OXC
Receiver Rx
Transmission fibre
Multiplexer MUX
Add-drop multiplexer ADM
Demultiplexer DEMUX
Fibre in N signals
ADM
Fibre out N signals
Drop signal
Add signal
(a)
λ_3
λ'_3
$\lambda'_1\lambda'_2\lambda'_3\lambda'_4$
Fibre 2 in N = 4 signals
λ_1
λ_2
λ_3
λ_4
Fibre 1 in N = 4 signals
OXC
Fibre 1 out N = 4 signals
λ_1
λ'_2
λ'_3
λ_4
Fibre 2 out N = 4 signals
(b)
$\lambda'_1\lambda_2\lambda_3\lambda'_4$

Figure B7.22. Schematic of a wavelength division multiplexed system. (*a*) Schematic of an optical add–drop multiplexer. As an example, there are four signals $\lambda_1 \cdots \lambda_4$ on the fibre. The signal λ_3 is dropped and replaced by λ'_3. (*b*) Optical cross connect with two input and two output fibres, each carrying four signals. In this example, signals 2 and 3 are exchanged between fibres 1 and 2.

  - ADMs (figure B7.22(*a*)) have $M = 2$ fibre ports. Their function is to locally drop and add a few wavelengths out of a stream of N signals carried on a fibre while directly passing on the other data streams. ADMs can be static or reconfigurable.
  - OXCs (figure B7.22(*b*)) have $M > 2$ aggregate ports and allow routing of any of the N signals from any input fibre to any output fibre.

  The hearts of reconfigurable ADMs and OXCs are *space switches*: for the ADM, an array of N 1×2 or 2×2 switches is required, and the OXC needs a matrix of dimension $N \times M$.

- At the end of the optical link, the signals are demultiplexed, and the *receiver Rx* converts the signals back into the electrical domain. The signal can be either terminated or regenerated; in the latter case, a fresh copy of the signal is produced through optical–electrical–optical (O–E–O) conversion and fed into the next link.

As a single fibre carries several gigabits/second of information, the system operators have to guarantee the reliability of the transmission system against failures such as a fibre break. *Network protection* techniques ensure reliability by providing redundant capacity in the network, and by using protection switches for re-routing the traffic in case of failures (figure B7.23).

Transmitter laser wavelength drifts and power variations are likely to cause an increase in BER, and it is therefore important for network operators to *monitor the performance* of the communication channels in order to guarantee quality of service (QoS) to their users.

Wavelength conversion of an optical signal is needed, for example, in OXCs if a signal gets switched onto a fibre in which a certain wavelength is already occupied.

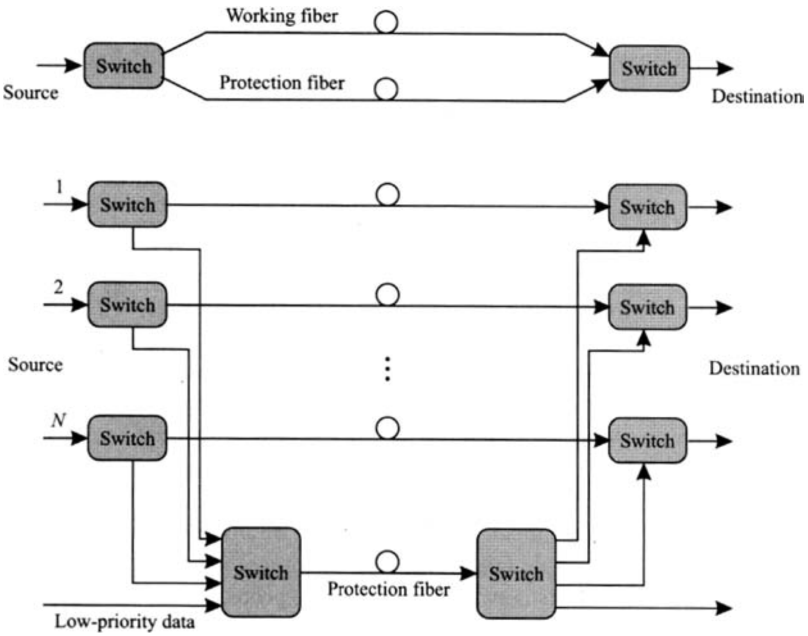


Figure B7.23. Protection in optical point–point networks by switching traffic from a working fibre on a protection fibre: 1:1 (top) and 1:N (bottom) protection [142]. The 1:N scheme is more bandwidth efficient as N working fibres share only one protection fibre.

Imperfect filters, dispersion, nonlinear effects and accumulated noise from Er^{3+} amplifiers limit the transparent length, over which a signal can be transmitted with a reasonable OSNR of about 1500–2000 km, and regeneration of the signal is required. As mentioned before, nowadays this is usually done at the receiver end in the electrical domain, but *regeneration in the optical domain* [107] has been demonstrated [108], resulting in extended reach [109]. Optical regeneration may eliminate a part of the cost of the current O–E–O conversion.

In current systems, a light path is set up for the duration of the communication (circuit switching). Future systems may evolve—in analogy to ATM or SONET packet switching in the electrical domain—to *optical packet switching* in which data packets are optically routed through the network based on information (such as the destination address) carried within the packet.

After the basic toolbox and technology overview in sections B7.2–B7.4, this last part will allude to examples of integrated devices. We will concentrate on some of the optical network functions mentioned above, and for avoiding duplication, we will not treat lasers, modulators, receivers and amplifiers for which we refer to chapters B1.2, B2, B4 and B6.

For most of the examples, several technologies are available and integrated waveguide devices represent one option, and the user will make his final choice based on performance, cost and maturity of the technology. Some of the functions are commercially available today, but most are still in research and will emerge over the next few years with the deployment of future generation systems.

B7.5.1 Optical multiplexers

Optical filters for WDM systems should have low insertion loss, low polarization and temperature dependence of both transmission wavelength and transmission loss, and they should not distort the incoming signals due to their average chromatic dispersion (nonlinear group delay of the filter) or dispersion variation within the filter bandwidth (dispersion slope) [110]. Technologies available today are mainly dielectric thin film filters, fibre Bragg gratings and planar devices.

Early wideband filters for separation of signals in the 1.3 μm/1.5 μm telecommunication windows have used single stage asymmetric Mach–Zehnder interferometers as described in chapter B7.2.4 [3], and wider passbands have been obtained by concatenation of three Mach–Zehnder filters [111]. Similar architectures are used for separating odd from even channels (interleaver) [112, 113].

Planar narrowband filters use the AWG (section B7.2.4) with typical channel spacings of 100 or 200 GHz (see also as an example, the 16 channel device in figure B7.7(*c*)). Commercial AWGs with Gaussian passband have fibre–fibre losses of the order of 4–5 dB, and in research devices with <1.2 dB insertion loss [114] and up to 1010 channels on a 10 GHz grid [115] (concatenated devices) have been demonstrated.

A wide transmission passband rather than a Gaussian passband is preferred for accommodating fluctuations in the transmitter laser Tx frequency and minimizing distortions due to filter concatenation [116]. Usual techniques modify the mode profile at the first AWG star coupler by MMI couplers [117] or Y-splitters [119], or interleave two gratings resulting in two separate focal points [120, 121]. Dynamic shaping of the passband is also possible [122].

The silica $\mathrm{d}n/\mathrm{d}T$ leads to a temperature dependence of the transmission spectrum, and for silica waveguide devices compensation methods [123, 124, 70] have been published. Polarization dependence has been eliminated by converting a TE mode into a TM mode in the middle of the device [125] and minimizing stress birefringence in the fabrication process [29, 30].

Both chromatic dispersion and dispersion slopes across filters also lead to signal degradation; however, the dispersion of the AWG filter is negligible and strictly zero if the component is symmetric and the loss can be neglected [118] (note that the transfer function in table B7.1 corresponds to that of an AR/FIR fitter without poles).

Other applications use very high resolution AWGs for converting a single data pulse from the frequency domain into the spatial domain. Examples are

- Shaping of ps pulses, which was reported first using dispersive bulk gratings and phase masks in the grating's focal plane [126] and later on using fully integrated AWGs with phase shifter arrays [2].
- Dispersion compensation [127] (section B7.5.3) and optical code division multiplexing [128].

B7.5.2 Dynamic gain equalization

Dynamic gain equalization is starting to be implemented in current systems, and a variety of technologies are being investigated today:

- Diffractive micro-electro-mechanical systems (MEMS) [129, 130].
- Bulk optic diffraction gratings in combination with liquid crystal spatial light modulators (LC-SLMs) [131, 132].
- Bulk and fibre acousto-optic gratings [133].
- Waveguides with surface electro-optic switchable Bragg gratings [134].

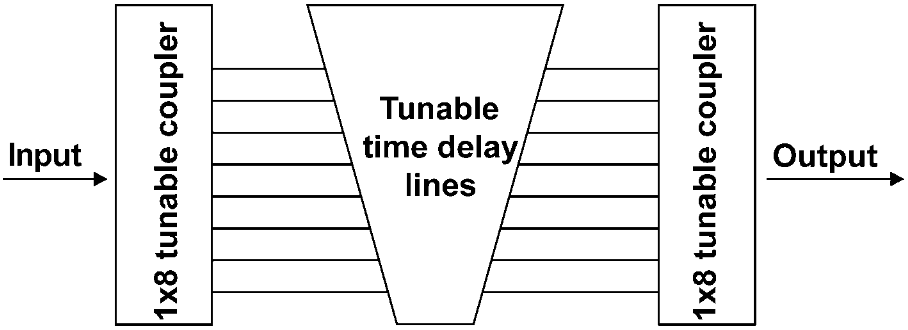
We will describe integrated optic solutions in more detail, and there are basically two types:

- Fourier filters—multiple sinusoidal transmission functions are generated by a multitude of asymmetric Mach–Zehnder interferometers (chapter B7.2.4) having path length differences ΔL. The ΔL results in different free spectral ranges, and superposition of the terms leads to the desired spectrum.

  The Mach–Zehnder devices can be arranged in parallel ('tapped delay line filter') or by concatenation ('serial lattice filter') [22, chapters 4.3 and 4.5].

  As an example, a schematic tapped delay line gain flattening filter [135] with eight arms (resulting in seven interference terms) is shown in figure B7.24 (top), and its transfer function is listed in table B7.1. Tailoring the delay T_k between adjacent arms and the amplitudes h_k allows synthesis of targets as in figure B7.24 (bottom). The typical insertion loss of such a device is 4 dB.

  Other applications of tapped delay lines are narrowband filters with small free spectral range [136] and multi-channel selectors [137]. Serial lattice gain flattening filters have been published using SiON waveguide technology [138], and silica waveguides [139].
- Channelized devices, which equalize individual wavelengths or groups of channels. An example is shown in figure B7.25 (top) [140]: a portion $\sqrt{R}$ of the light propagates through the upper unfiltered arm, and $\sqrt{(1-R)}$ will go through the lower arm. This lower arm contains a demultiplexer for selecting a single wavelength or a band of wavelengths and an array of phase shifters. Depending on the setting of the phase shifters, constructive or destructive interference between upper and lower arms results for each spectral band. Devices have been fabricated in both InP [140] and silica waveguides [141], and for the latter device a typical spectrum is shown in figure B7.25 (bottom). The insertion loss through the device is around 6.5 dB.



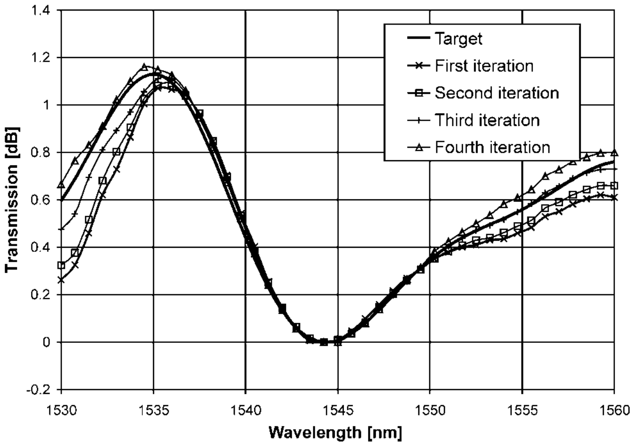


Figure B7.24. Tapped delay line filter for dynamic gain flattening. The device consists of a tunable 1 × 8 splitter made from cascaded MZIs, the tunable delay lines and an 8 × 1 combiner (top) architecture. Feedback is given to an electronic control circuit, and the measured transmission after four iterations (triangles) is closer than 0.1 dB to the target (thick solid line) [135]. The transmission is normalized to 0 dB. Reproduced by permission of IEEE.

B7.5.3 Chromatic and polarization mode dispersion compensation

Spreading of the light pulses propagating through a fibre arises from chromatic dispersion of the fibre itself, from nonlinear effects leading to a modulation of the phase, and from the chromatic dispersion of filters present in the link. As a consequence, the pulses in adjacent bit periods will start to overlap (intersymbol interference (ISI)) leading to an increased BER.

For a signal with bit rate B and a fibre of length L (km) with dispersion D ($\text{ps}\,\text{nm}^{-1}\,\text{km}^{-1}$), the condition

$$B^2LD < 11\,000\,\text{ps nm}^{-1}(\text{Gbit/s})^{-2}$$

has to be fulfilled for a PP of <1 dB [142]. In other words, a 10 Gbit s^{-1} signal tolerates an accumulated dispersion LD of 1100 ps nm^{-1}, but for a 40 Gbit s^{-1} signal, the limit is only 69 ps nm^{-1}.

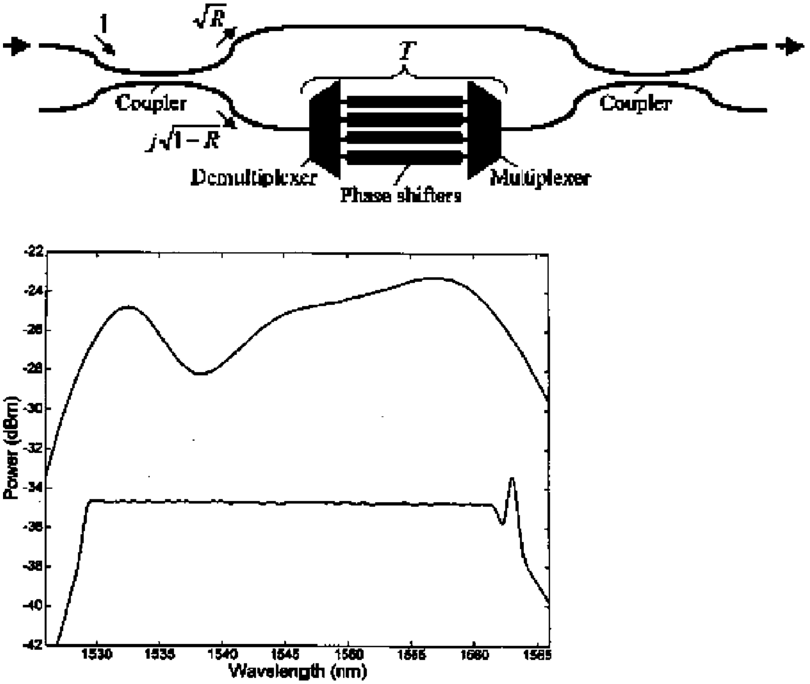


Figure B7.25. Channelized dynamic gain flattening filter [140, 141]—device architecture (top) and result of automatic flattening of the emission from an erbium doped amplifier (bottom), the upper and lower curves corresponding to before and after flattening. Reproduced by permission of IEEE.

In practice, the chromatic dispersion of standard single mode fibre [82] $D = +17\,\mathrm{ps\,(nm\,km)^{-1}}$ is compensated by periodically adding a section of dispersion compensating fibre (DCF) having negative dispersion, thus bringing the cumulated dispersion back to tolerable values. However, the tight limits for $40\,\mathrm{Gbit\,s^{-1}}$ signals and variations in the fibre dispersion (through fabrication tolerances and daily temperature changes) may necessitate tunable dispersion compensation devices operating on single channels or groups of channels.

In addition to the dispersion D, the dispersion slope $\mathrm{d}D/\mathrm{d}\lambda$ across a single WDM channel needs to be compensated for ultra-high speed systems ($>100\,\mathrm{Gbit\,s^{-1}}$) as well.

Technologies for dispersion compensation are chirped fibre Bragg gratings [143], micro-optic Gires–Tournois interferometers [144] and virtual phased arrays [145], and in the following we will describe the integrated optic devices, which offer an opportunity for integration with the multiplexers at the end of the transmission link.

Planar dispersion compensators have been studied based on both Mach–Zehnder filters and ring resonators, corresponding to the FIR and IIR categories (section B7.2.5).

An FIR dispersion slope compensator is shown in figure B7.26 (top) [146], and in this case integration with an AWG multiplexer allows us to compensate 16 WDM channels simultaneously.

For each channel, the dispersion equalizer consists of five asymmetric Mach–Zehnder filters and six tunable couplers. The asymmetric MZIs provide a wavelength sensitive splitting and a delay between the signals propagating through the upper and lower arms, resulting in a wavelength dependent delay. The device itself is fabricated in $1.5\%\Delta$ silica waveguides, and the thermo-optic phase shifters allow tuning of

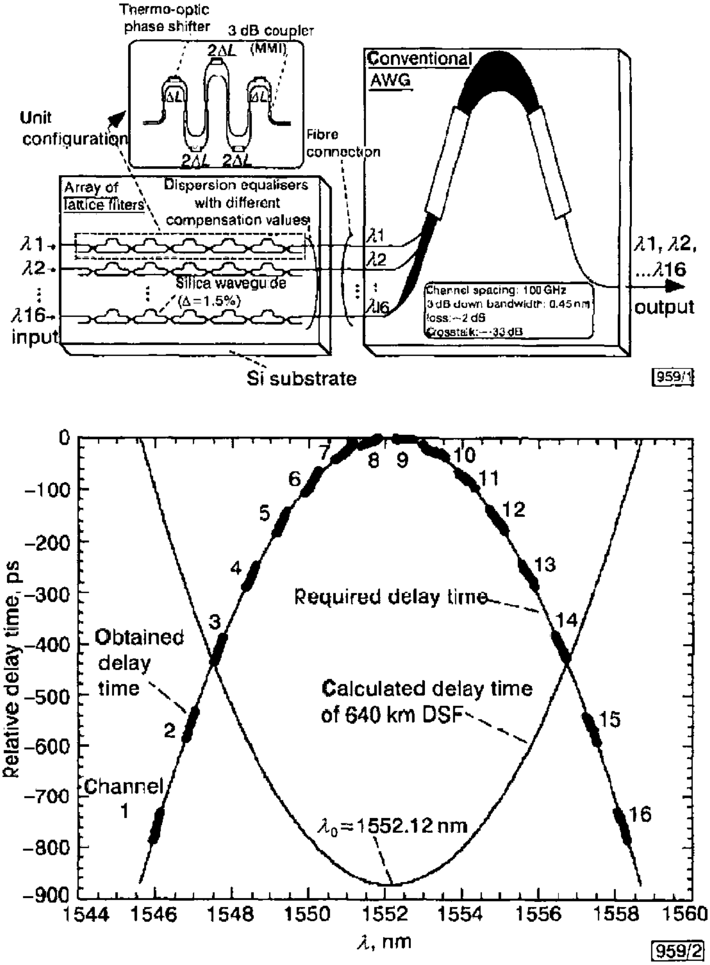


Figure B7.26. Sixteen channel dispersion slope compensating lattice filter (top). The dispersion slope is compensated separately for each wavelength through a five stage lattice filter. The bottom figure shows the delay of the transmission fibre, the required delay and the experimental delay (full symbols) [146]. Reproduced by permission of IEEE.

the dispersion characteristics. The measured versus required delay for transmission over 640 km dispersion shifted fibre (DSF) is shown in figure B7.26 (bottom).

An alternative architecture for compensation of both dispersion and dispersion slope is based on a pair of AWGs and a filter (figure B7.27)[127]: the AWG separates the incoming signal into its frequency components, and the spatial phase filter will apply the appropriate delay. Designs for compensation of up to 260 ps nm^{-1} are proposed, but losses are expected to be of the order of 15 dB due to the specific design of the AWG and coupling losses from the AWG to the spatial phase filter.

IIR filters are particularly interesting, since the poles in their transfer function lead to a nonlinear phase response and thus naturally provide dispersion.

In particular, the example shown in figure B7.28 (top, see also the transfer function in table B7.1) [147] represents a two stage all-pass filter which is theoretically lossless, and the dispersion can be

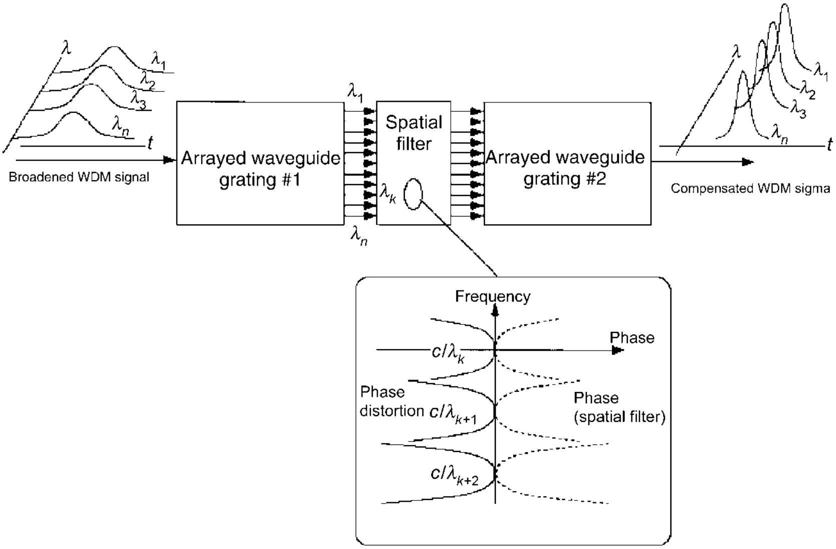


Figure B7.27. Dispersion compensator using arrayed waveguide gratings [127]. AWG 1 will spread the spectrum, and for each WDM channel there is a spatial filter (shown as an inset) which will correct for the phase distortion/ chromatic dispersion. AWG 2 then recombines the compensated signals. Reproduced by permission of IEEE.

tailored by adjusting the power coupling ratios κ_1, κ_2 and the phase shifts ϕ_1, ϕ_2. A dispersion of $\pm 4000\,\mathrm{ps\,nm^{-1}}$ has been demonstrated over a bandwidth of 4.5 GHz (figure B7.28, bottom), and more recently [148] the bandwidth has been increased to 13.8 GHz (dispersion $\pm$ 2000 $\mathrm{ps\,nm^{-1}}$) by using a silica waveguide fabrication process with a $2\%\Delta$ index contrast and thus shorter feedback path.

PMD arises from ellipticity of the transmission fibre as well as from polarization dependent performance of the components within the transmission link, and as a result TE and TM modes experience a differential delay. PMD is a time dependent statistical phenomenon and needs to be compensated adaptively.

Early PMD compensators [149] were based on a series of three squeezed polarization maintaining fibres, and a planar solution is described in [150] (figure B7.29, top). The compensator splits the incoming signal into TE and TM components at the first polarization beam splitter (using stress birefringence) [151], then TM is converted into TE allowing interference between TE and TM modes in the subsequent serial MZI. In this particular case, the group delay difference for the two polarization states is compensated in a fixed 7.5 ps delay line, and then the polarization modes are recombined. The systems benefit is demonstrated in figure B7.29 (bottom), showing the BER for a transmission link with (□) and without compensator (Δ).

B7.5.4 *Space switches and all-optical switching*

In the beginning of this section we mentioned multiple applications for switches in an optical network:

- Space switches in OXCs and add–drop multiplexers enable dynamic provisioning or reconfiguration of an optical WDM lightpath (circuit switching), thus replacing manual fibre patch panels. For OXCs, the number of input/output ports needed is difficult to estimate, but the many

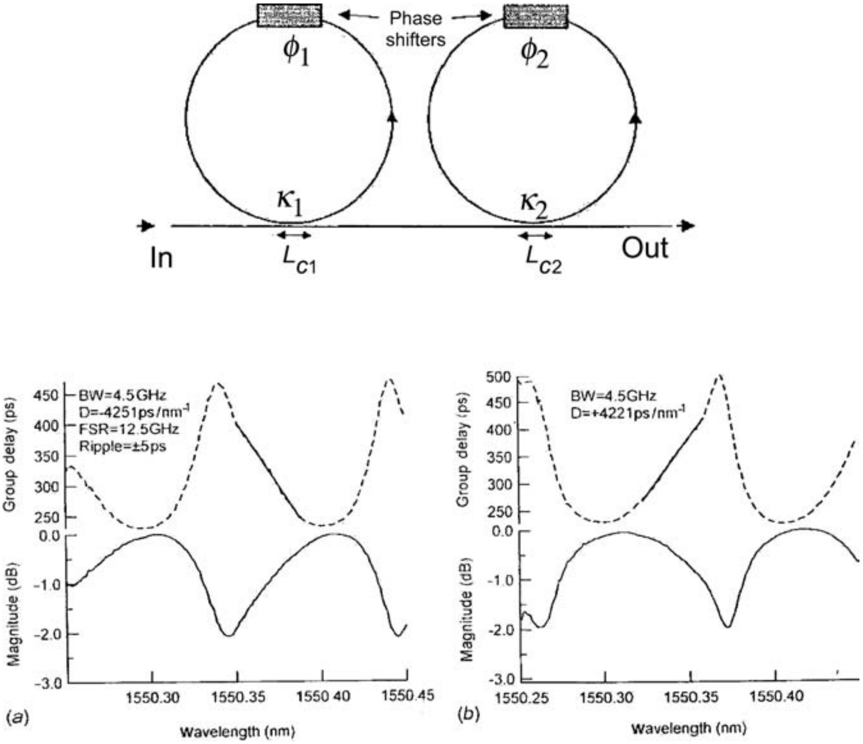


Figure B7.28. Two stage optical AR filter based on ring resonators (top) [147] for compensation of both positive and negative chromatic dispersion. Each stage has a feedback path with radius $R = 2.2$ mm, resulting in a free spectral range of about 0.12 nm (15 GHz). Measured group delay and normalized insertion loss are shown for negative (bottom, left) and positive (bottom, right) dispersion. Reproduced by permission of the Optical Society of America.

wavelengths carried on one fibre (at present usually $N = 40-80$) and the large number of fibres M will drive the demand for a high port count (256×256 and larger) devices. The switching speed required for this application is of the order of 1–10 ms.

- Fibre optic networks can encounter two types of failure: problems with the transmission equipment or path interruptions due to fibre break. Both issues can be addressed by switching the traffic onto a second unused fibre which is dedicated (1:1 protection) or shared with other traffic streams (figure B7.23). Protection switches are required to have a commutation time of <10 ms.
- Future systems will use high speed switches for switching and routing of data packets rather than provisioning a dedicated lightpath. The required switching speed is inversely proportional to the bit rate, i.e. it is of the order of a few nanoseconds.

Electronic data processing seems to be currently limited to around 40 Gbit s^{-1}, and there is an interest performing switching (and more generally signal processing such as multiplexing, routing, wavelength conversion, optical logic) in the all-optical domain [152]. As an example, we will briefly mention all-optical switches at the end of this section, and we refer to chapter C1.3 and reference [63, chapter 9] for all-optical time > 1 Tbit s^{-1} systems.

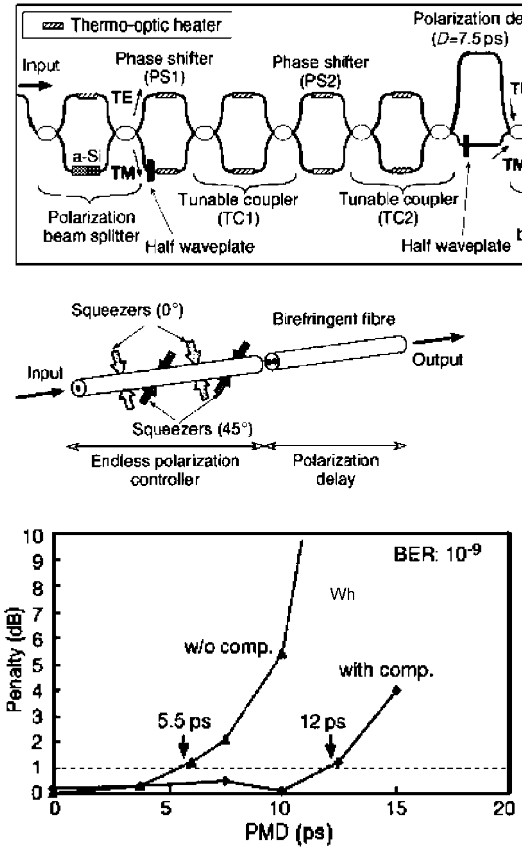
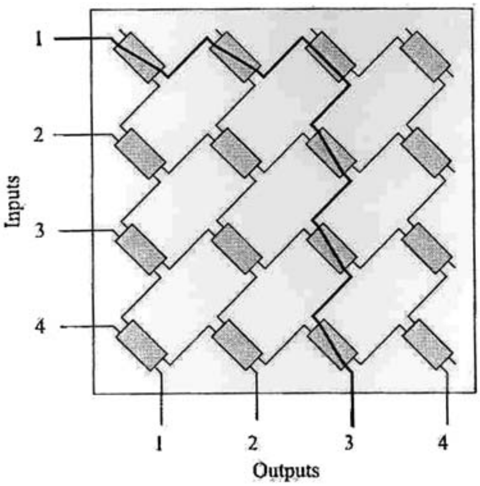

Thermo-optic heater
Polarization delay (D=7.5ps)
Input
Phase shifter (PS1)
Phase shifter (PS2)
Output
TE
TM
a-Si
Tunable coupler (TC1)
Tunable coupler (TC2)
Polarization beam splitter
Half waveplate
Half waveplate
Polarization beam splitter
Squeezers (0°)
Birefringent fibre
Input
Output
Squeezers (45°)
Endless polarization controller
Polarization delay
BER: 10^-9
Wh
w/o comp.
with comp.
5.5 ps
12 ps
Penalty (dB)
PMD (ps)
0
1
2
3
4
5
6
7
8
9
10
0
5
10
15
20

Figure B7.29. Integrated planar lightwave circuit PMD compensator and equivalent circuit [150] (top). Measured power penalty as a function of the polarization mode dispersion for a fixed bit error rate (BER) of 10^{-9}. Reproduced by permission of IEEE.

In addition to the commutation time, other important switch characteristics are extinction (or ON–OFF) ratio, insertion loss, crosstalk (i.e. leaking of light from one path into another), polarization dependent loss and power consumption.

The basic elements are 1×2 or 2×2 switches, and larger size devices can be realized by appropriate cascading of these devices. As an example out of many possible arrangements, the crossbar and the Beneš architectures [142] are shown in figure B7.30, and the choice between architectures will depend on criteria such as

- The number of 1×2s or 2×2s. An $n \times n$ crossbar switch consists of n^2 elementary switches, and Beneš switches need only $n/2{*}(2{*}\log_2(n) - 1)$.
- Loss and loss uniformity—in the Beneš switches, each signal crosses the same number of 2×2s, but in the crossbar switches there is a shortest path (going through one switch only) and a longest path (through $2n - 1$ elements).



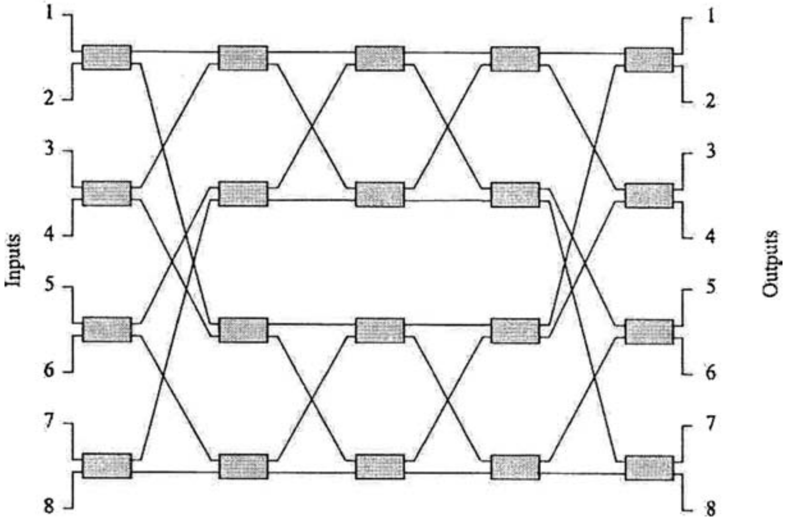


Figure B7.30. 4 × 4 switch: crossbar (top) and Beneš architectures (bottom) [142].

- The number of waveguide crossovers—this will lead to crosstalk between the lightpaths. Crossbar switches have no crossovers, in contrast to Beneš switches.
- Switches can be blocking or nonblocking. In the latter, a connection can be made between any unused input port and any unused output port. If this new connection can be made without interrupting the traffic on the other channels, the switch is called 'wide sense nonblocking' (the crossbar switch belongs to that category); otherwise it is called 'rearrangeably nonblocking' (e.g. Beneš switches).

Other architectures are described in [153, 142 chapter 3.7].

The elementary 1 × 2 or 2 × 2 optical switches themselves use a wide range of operating principles, which can be roughly classified into:

		Operating principle	
		Interferometric	Noninterferometric
Switching mechanism	Mechanical		Mechanical micro-optic systems, MEMS
	Nonmechanical	Mach–Zehnder interferometer: thermo-optic effect in doped silica and polymer waveguides, electro-optic LN switches	Y-branch devices with polymer thermo-optic, LN and InP electro-optic effects
		Hybrid silica/polymer tunable vertical coupler switch	Total internal reflection: silica–air interfaces, InP electro-optic
		All-optical $\chi(3)$ switches in GaAs, InP	Polarization switching in liquid crystals
			Semiconductor optical amplifier (SOA) ON–OFF gate

Mechanical switches route light through a physical displacement of the light path: the fundamental switching elements are movable mirrors or prisms, which are mechanically actuated. Another approach is to steer beams from N input fibres on N output fibres, the difficulty being that an analogue control of the beam position is required.

Main advantages of mechanical switches are low insertion losses and high crosstalk between channels. However, their commutation speed is of the order of tens of milliseconds to seconds, and their size is limited to about 32 × 32.

An emerging technology is MEMS, which combine the optical performance of mechanical switches with the benefits of integrated optics (compactness, low cost volume production, optical prealignment) [154, 155, 156]. Schematic drawings and SEM pictures of the two-dimensional (2D) matrix and the movable mirrors are shown in figure B7.31. The mirrors are fabricated via Si surface micro-machining; the translation movement of a scratch drive is translated into a rotation, and as a result the mirror flips between the two states. For 8 × 8 2D devices insertion losses of 3.1–3.5 dB with <1 ms switching speed have been reported, and large scale cross connects with <6 dB losses and 512 × 512 ports or higher are expected to be constructed out of 32 × 32 modules [155].

Higher port count monolithic devices are accessible through MEMS mirrors with two rotational degrees of freedom. The continuous rotation is used for steering the beam in all three dimensions, and the MEMS panel is placed between an array of input/output fibres and a fixed reflecting mirror. Switches with 112 × 112 ports have been demonstrated [157].

In general, mechanical switches use expanded beam optics to minimize coupling losses.

Nonmechanical noninterferometric switches depend on the control of waveguide index or polarization state.

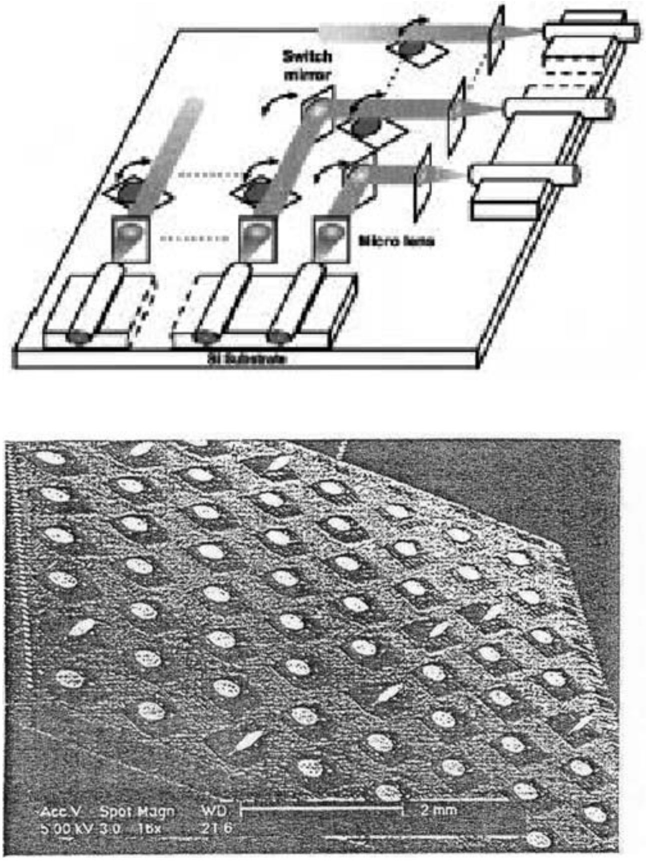


Figure B7.31. Free space MEMS switch: schematic [200] and scanning electron microscope picture [155] of a switch matrix (above). Schematic drawing of a single mirror and SEM close-up (next page). Reproduced by permission of IEEE and the Optical Society of America.

In a fixed symmetric Y-branch (section B7.2.3), light is equally divided between the output waveguides, but dynamic control of the refractive indices allows us to ‘push’ or ‘pull’ light into one or the other output.

Layout and switching characteristics for a 1 × 2 digital optical polymer waveguide switch are shown in figure B7.32 [7]: increasing the temperature of the upper arm will decrease the waveguide index, thus coupling more light into the lower arm. Typical power consumption is around 50–80 mW with about 1 ms commutation time.

Electro-optic switches can use the same principle, but provide higher speeds. Hundreds of megahertz have been reached using Ti:$LiNbO_3$ waveguides [158], and crosstalk has been enhanced to 30–50 dB by concatening 1 × 2 switches (‘dilated switch’). Although power consumption is low, the main inconvenience is the applied voltage of about 25 V. Monolithic 4 × 8 and 6 × 6 LN switches are reported in [159].

Crossconnects based on EO-InP switches combined with SOAs are particularly promising for nanosecond packet switching applications, as they provide high extinction ratio and can be lossless. Figure B7.33(*a*) shows an elementary switching element [160]: in the Y-junction the switching is done by

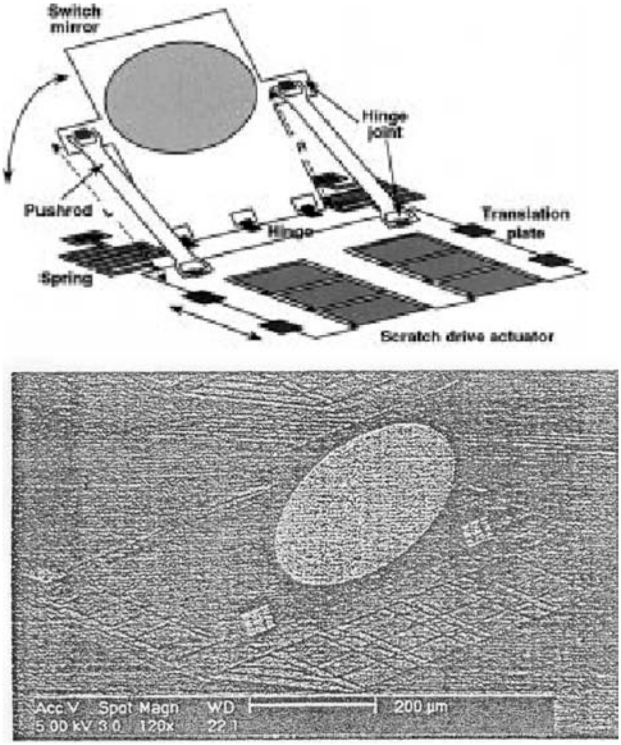


Figure B7.31. (Continued)

current injection and the associated index change (section B7.2.2). The travelling wave amplifier TWA compensates for loss when the switch is in the ‘ON’ state and suppresses crosstalk from other channels when the switch is ‘OFF’. Fibre-to-fibre performance of a 4 × 4 crossbar switch operating at 1.31 µm wavelength is depicted in figure B7.33(*b*)—the device is lossless for an injected current of >200 mA, and the extinction ratio is in excess of 40 dB. Polarization sensitivity is an issue but can be reduced to <1 dB by using square shaped waveguide cross-sections [161].

Switches using total internal reflection (TIR) at silica waveguide–air interfaces are reported in [162], [163] and [164]. The principle is shown in figure B7.34 (top): a drop of index matching fluid can be moved within a narrow trench, leading to transmission or TIR.

An example of asymmetric Y-couplers and electro-optic induced TIR is shown in figure B7.34 (bottom) [165].

Polarization rotation in liquid crystal (LC) spatial light modulators in conjunction with birefringent crystals results in space switching. The first birefringent polarization beam splitter in figure B7.35 (bottom) [166] will separate the beam into ordinary and extraordinary beams having orthogonal polarizations, and the arrayed $\lambda/2$ plate symbolizes a matrix of liquid crystal modulators. Depending on the orientation of the LC molecules (which in turn can be changed by application of an electric field) the polarization of incoming light is rotated by 90° or remains unchanged. The subsequent birefringent crystals will recombine ordinary and extraordinary rays.

SOAs can be used as a gate (ON/OFF switch) by the variation of the applied bias voltage: a low voltage results in low population inversion, and an incoming signal gets absorbed. Application of the bias voltage leads to an increase in population inversion, resulting in the transmission and gain of

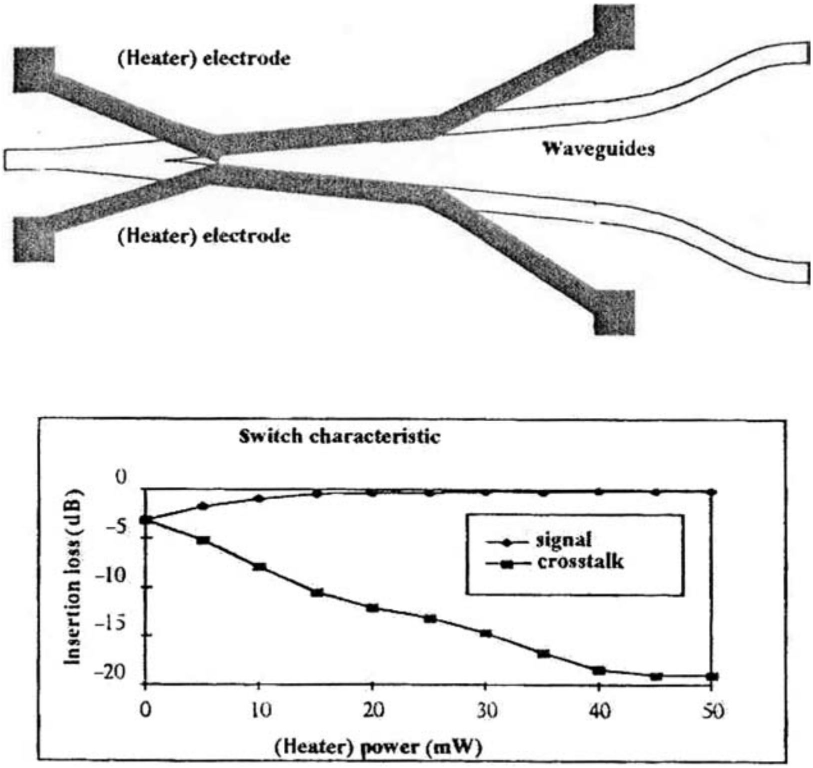


Figure B7.32. 1 × 2 digital optical Y-branch switch: layout with waveguides and heating electrodes for TO effect, extinction as a function of the applied power [7]. Reproduced by permission of Marcel Dekker Inc.

the signal. The commutation speed is of the order of 1 ns [142]. 1 × 4 [167], 4 × 4 [168] and 8 × 8 space switches [169] operating at 10 Gbit s^{-1} have been demonstrated based on a broadcast-and-select (the incoming signal is split, then the SOA ON/OFF gates pass through or block the signals) architecture. Devices have been fabricated through hybrid integration of the SOA gates on SiO_2 motherboards.

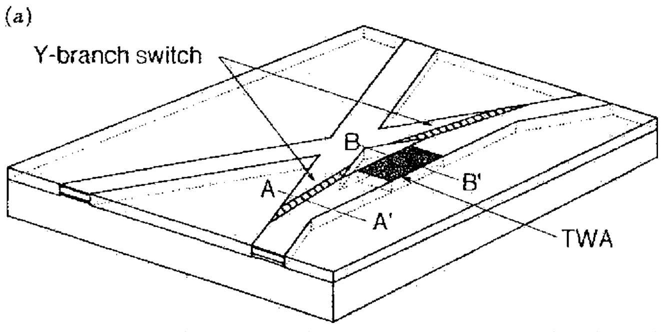


Figure B7.33. (*a*) Elementary semiconductor switch element consisting of Y-branches and a semiconductor travelling wave amplifier for loss compensation. (*b*) Performance of a 4 × 4 switch: the device is lossless for >200 mA injection current, and the extinction ratio is >40 dB [160]. Reproduced by permission of World Scientific Publishing Co Ltd.

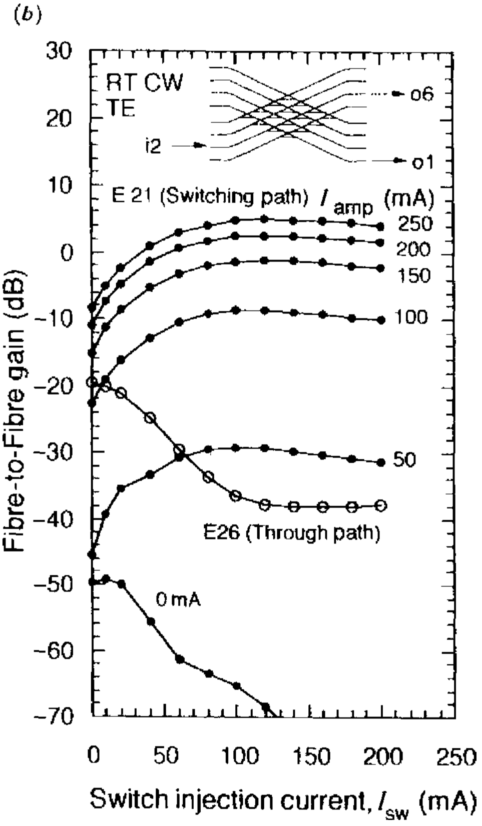


Figure B7.33. (Continued)

Another application for SOA gates is to select a set of wavelengths from a WDM signal, and both hybrid [170] and monolithic integrated InP [171] devices have been demonstrated (figure B7.36): the incoming WDM signal is wavelength demultiplexed in a first arrayed waveguide grating, passed through or blocked in the SOAs and then recombined in a second arrayed waveguide grating.

Nonmechanical interferometric switches are based on tunable directional couplers or Mach–Zehnder interferometers (sections B7.2.3 and B7.2.4) and used in thermo-optic, linear (Pockels) or nonlinear (Kerr) electro-optic effects for achieving the required phase shift.

Monolithic 16 × 16 crossbar switch matrices using silica waveguide MZIs and the thermo-optic effect fabricated on a 6 inch substrate are described in [172] (figure B7.37). The high extinction ratio of 55 dB is achieved by concatenation of two MZIs ('dilated switch'), and the average insertion loss is 6.6 dB. The electrical power consumption per switch point is 450 mW resulting in a total of 17 W. However, in recent 2 × 2 and 8 × 8 switches a reduction to 45 mW per stage has been demonstrated by micro-machining grooves on either side of the waveguide and insertion of a thick silica buffer layer for enhanced thermal isolation between the waveguide and the Si substrate [173, 174].

Polymer waveguides and their large dn/dT further reduce the power consumption, and for a 2 × 2 device, 5 mW [175] have been reported in the 1.3 μm telecom window.

The vertical coupler 1 × 2 switch shown in figure B7.38 [67] combines the low propagation loss in silica with the low switching power of polymer devices: the structure consists of a vertically

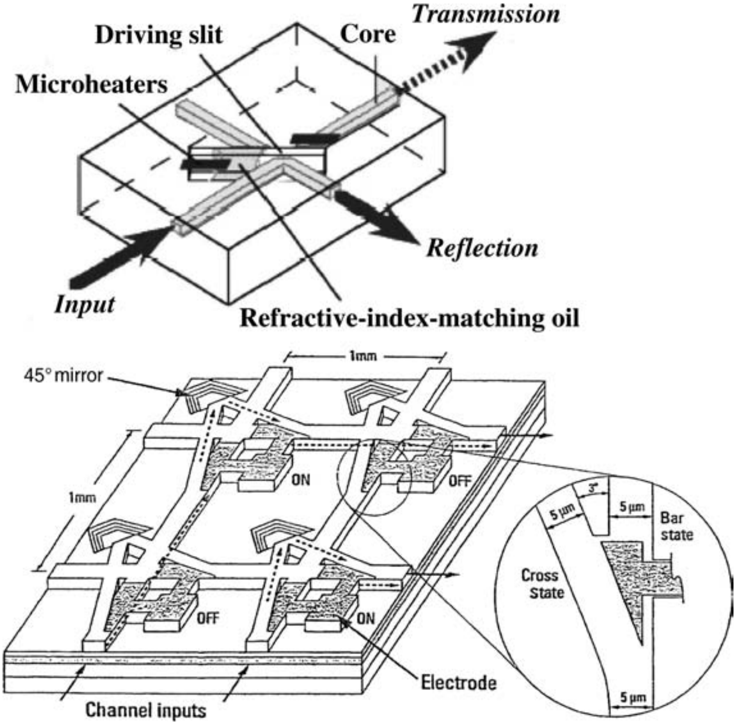


Figure B7.34. Total internal reflection switch with silica waveguides (top, [163]). The left part of the driving slit is filled with index matching oil. Application of heat and capillarity will drive the oil into the intersection of waveguides and switches from reflection to transmission. TIR EO switch (bottom) [165]. Reproduced by permission of the Optical Society of America and the American Institute of Physics.

stacked lower doped silica waveguide layer, a thin silica cladding and the polymer waveguides. When the electrodes are not powered (OFF), input light coupled to the lower silica waveguide will propagate through this silica waveguide (bar state). When the electrodes are sufficiently powered (ON), the input light is coupled to the polymer waveguide in the first element and coupled back to another parallel silica waveguide in the second element (cross state). The reported switching power was <80 mW and crosstalk <20 dB.

For electro-optic MZI $LiNbO_3$ waveguide switches see reference [176].

Forward current injection into InGaAsP/In structures leads to a negative index change of $-0.4\,\mathrm{rad\,mA^{-1}}$ [177], and application of a reverse voltage [178] increases the refractive index, leading to a positive phase change of $0.25\,\mathrm{rad\,V^{-1}}$.

All-optical switching is a very attractive concept for ultra-high speed systems ($>100\,\mathrm{Gbit\,s^{-1}}$) as bandwidth-limited electronic processing is replaced by the manipulation of an optical data signal via a second optical control pulse [7, chapter 11.3] (see chapters A2.4 and B6).

Third order nonlinearity $\chi(3)$ (section B7.2.2) in GaAs has been used in the monolithic device shown in figure B7.39 (top) [179]: when the first control pulse is absorbed in the upper nonlinear waveguide the Mach–Zehnder interferometer becomes asymmetric and the incoming signal is switched from one output to the other. The second control pulse then cancels the index change from the first

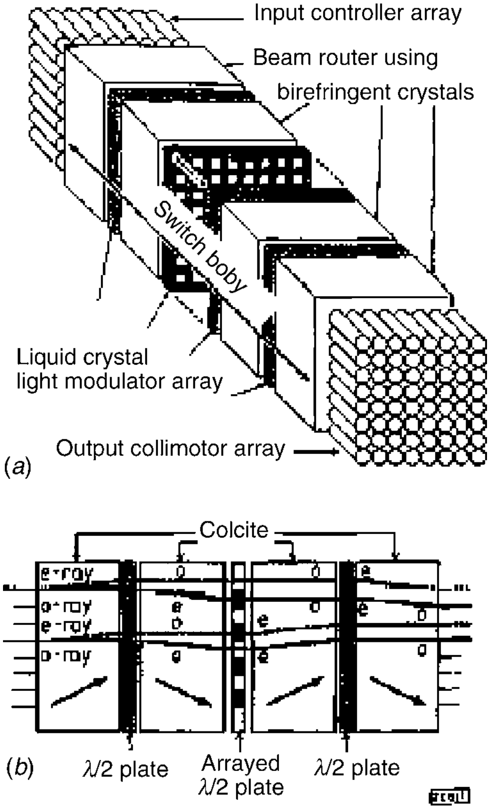

Input controller array
Beam router using
birefringent crystals
Switch boby
Liquid crystal
light modulator array
Output collimotor array
(a)
Colcite
e-ray
o-ray
e-ray
o-ray
(b)
λ/2 plate
Arrayed
λ/2 plate
λ/2 plate

Figure B7.35. Liquid crystal free space optical switch (top) with input/output collimator arrays and polarization based switching (bottom) [166]: the incoming beam is split into o and eo polarizations. The liquid crystals represent a switchable array of halfwave plate, i.e. they turn (or not) the polarization by 90°. The right part is symmetric to the left part. Reproduced with permission from IEEE.

pulse, thus switching the signal back. The control pulses with an energy of <7 pJ and a length of 1.3 ps came from a Ti:sapphire laser, and switching speeds were 8 ps (figure B7.39, bottom).

B7.5.5 Integrated add–drop multiplexers and optical crossconnects

Add–drop multiplexers are wavelength selective switches allowing dropping of one or more signals from an incoming data stream and adding signals to the outgoing data stream. Different monolithic arrangements based on AWGs (chapter B7.2.4) and switches are shown in figure B7.40 (top) [17], and configurations (*a*) and (*b*) use the same AWG for demultiplexing and multiplexing the incoming traffic into individual wavelength channels. The loop back configuration (*a*) has been chosen for the monolithic InP component in figure B7.40 (bottom) [180], which is designed for four channels on a 200 GHz grid. The component is extremely compact ($3 \times 6\,\mathrm{mm}^2$) and uses electro-optic MZI 2 × 2 switches for passing through or adding/dropping traffic on each of the four channels.

Another device architecture for ADM is the serial lattice filter (chapter B7.5.3) made of SiON waveguides (figure B7.41, top) [144]: there are 12 asymmetric Mach–Zehnder stages resulting in a chip size of $6 \times 65\,\mathrm{mm}^2$. The device operates on a 200 GHz grid and can dynamically add/drop one

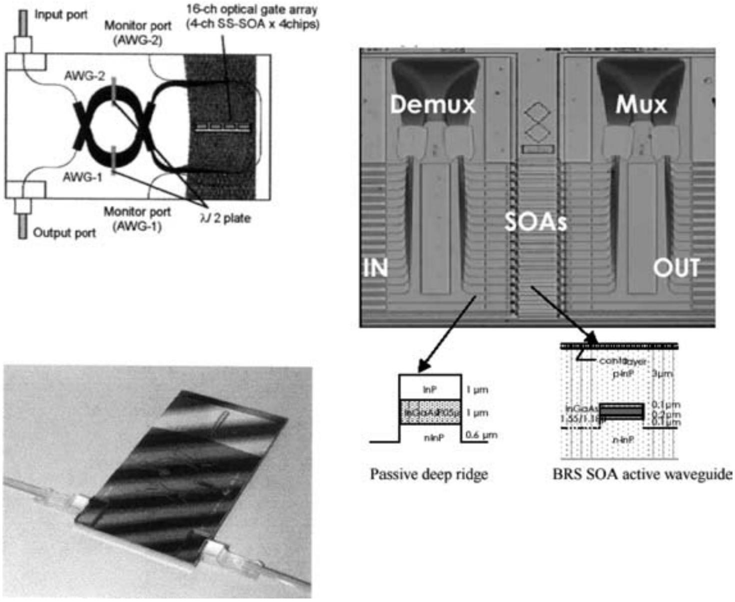


Figure B7.36. Hybrid (left, [173]) and monolithic InP (right, [171]) channel selector. In both cases, a DEMUX will separate the wavelengths, which are gated through InP SOAs. The second MUX then re-combines the channels. Reproduced by permission of IEEE.

out of the incoming eight channels. The 12 MZI stages give sufficient finesse for suppressing crosstalk from adjacent channels as shown in the theoretical filter response (figure B7.41, bottom). Experimental on-chip losses at 1550 nm were about 2 dB, isolation of the through channels from the dropped wavelengths was around 20 dB and tuning was accomplished with thermo-optic phase shifters.

Polymer waveguides with their 10 × higher thermo-optic coefficient $\mathrm{d}n/\mathrm{d}T$ and the possibility of writing Bragg gratings (see chapter B7.3.5) provide another possible architecture when combined with four-port circulators (figure B7.42) [64]. The function of the Bragg grating is to separate the drop wavelength by reflecting it back to the left circulator and to pass all other wavelengths through. The main challenge is the fabrication of the Bragg grating with uniform and strong reflection within a narrow wavelength band, no out-of-band spectral features, low insertion loss and polarization dependence.

Other material systems for AD filters are Si waveguide ring resonators [181] and acousto-optic $\mathrm{LiNbO_3}$ Mach–Zehnder switches [182, 63 chapter 7.8].

The integrated 2 × 2 optical cross connect shown in figure B7.43 [183] consists of AWGs giving simultaneous access to all 16 wavelengths and an array of thermo-optic 2 × 2 switches: light entering AWG1 is demultiplexed and enters the left hand side of the switches. When the switches are ‘OFF’, the signals will be recombined in AWG3; otherwise they are routed to AWG4. The second data stream coming into the OXC/AWG2 will arrive at the right hand side of the TO switches and go through to AWG4 (‘OFF’) or is exchanged to AWG3 (‘ON’). The switches consist of two stage MZIs for reaching an extinction ratio of <-28 dB. A transmission spectrum is shown in figure B7.43 (bottom). The size of the device was $87 \times 74\,\mathrm{mm}^2$ and insertion losses between 7.8 and 10.3 dB.

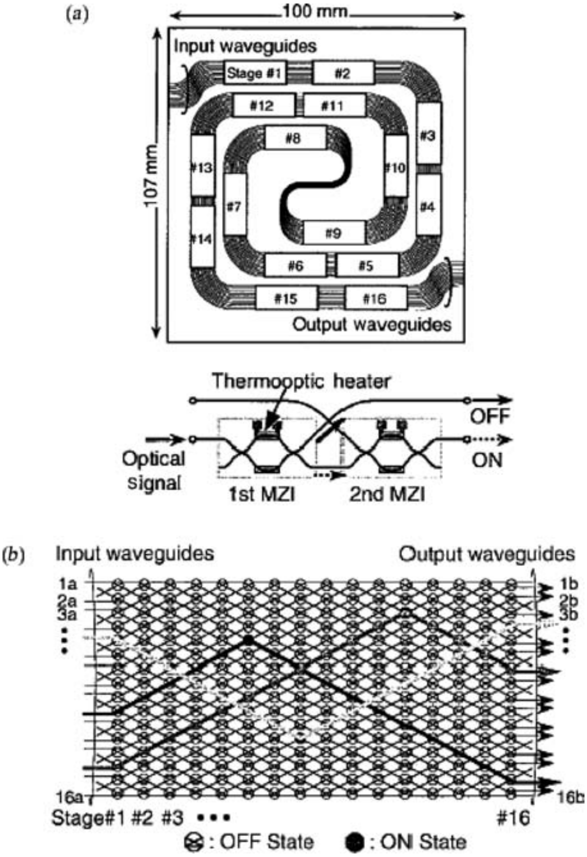

(a)
100 mm
107 mm
Input waveguides
Stage #1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
#13
#14
#15
#16
Output waveguides
Thermooptic heater
Optical signal
1st MZI
2nd MZI
OFF
ON
(b)
Input waveguides
Output waveguides
1a
2a
3a
16a
1b
2b
3b
16b
Stage#1 #2 #3 • • •
#16
⊗ : OFF State
● : ON State

Figure B7.37. Silica waveguide thermo-optic switch (16 × 16) [172]: layout on the 6 inch wafer (top) and architecture (bottom). Each switch point consists of a two stage Mach–Zehnder for enhanced ON–OFF extinction ratio. Reproduced by permission of IEEE.

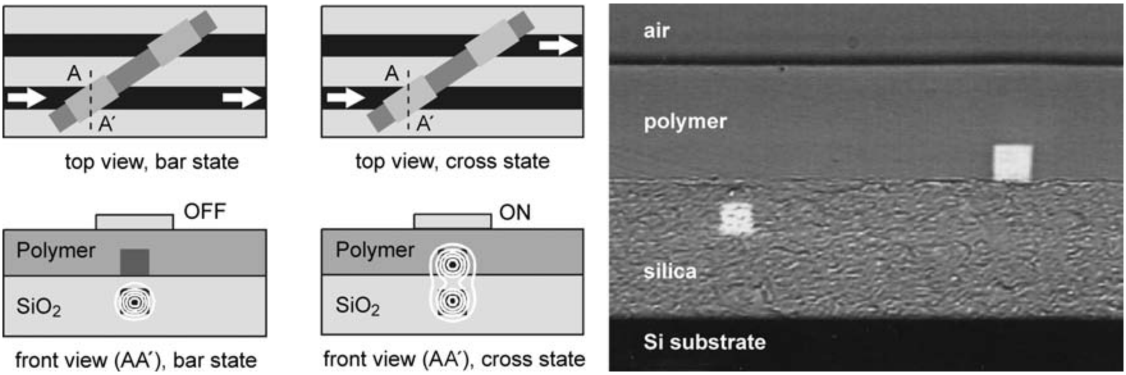

A
A′
top view, bar state
top view, cross state
OFF
ON
Polymer
SiO2
front view (AA′), bar state
front view (AA′), cross state
air
polymer
silica
Si substrate

Figure B7.38. Vertical coupler switch (VCS) [67]: schematic layout of the 1 × 2 hybrid polymer/silica VCS and calculated field distribution (left); cross-section of a fabricated vertical coupled structure (right). Reproduced by permission of VDE Verlag.

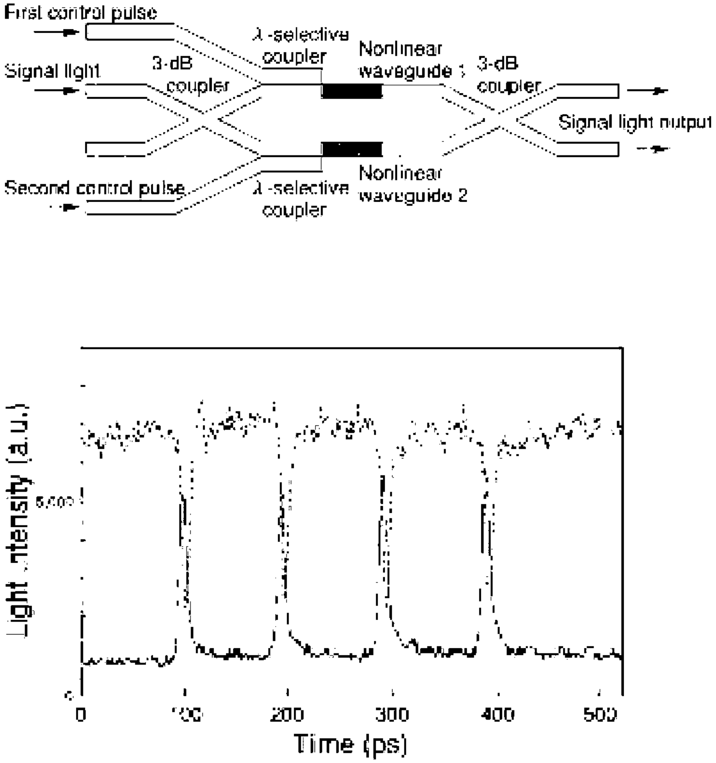

First control pulse
Signal light
3-dB coupler
λ-selective coupler
Nonlinear waveguide 1
3-dB coupler
Signal light output
Second control pulse
λ-selective coupler
Nonlinear waveguide 2
Light intensity (a.u.)
5,000
0
0
100
200
300
400
500
Time (ps)

Figure B7.39. Monolithic all-optical symmetric Mach–Zehnder switch using nonlinearity in GaAs [179]. Device layout with inputs for the signal and the two control pulses defining a gate (top); switching characteristics (bottom). Reproduced by permission of the American Institute of Physics.

A simpler arrangement (figure B7.44, top) [184] uses only two AWGs. The AWGs have an interleaved chirped grating, thus producing two separate images for each wavelength in different grating orders (denoted by ω_0, ω_{-1}, ω_{+1}). Per channel, there are three connecting waveguides with InP phase shifters between the two AWGs, allowing the dynamic re-routing of wavelength channels to either line 1 or line 2 of the second AWG. The footprint of the device is $4.2 \times 9.8\,\mathrm{mm}^2$ [177], and experimental on-chip spectra for a 2×2 crossconnect with six channels are shown in figure B7.44 (bottom). Reported on-chip losses were about 10 and 19 dB including fibre/chip coupling.

B7.5.6 Integrated monitoring devices

Among the different network management functions [142 chapter 10.1], performance monitoring deals with providing a guaranteed QoS to the network users. As an example, a drift of the transmitter laser Tx wavelength due to temperature variations or ageing can lead to a decrease in received signal power (and thus increased BER), when the laser wavelength is not centred any longer on the passbands of the MUX/DEMUX filters. The silica waveguide component in figure B7.45 [185] simultaneously monitors 16 WDM channels present in a system. The operation principle is to feed a WDM signal (the wavelength is marked by the circle in figure B7.45, bottom) into an arrayed waveguide grating with large passbands and to compare the received power on adjacent outputs. Misalignment of the laser frequency will result

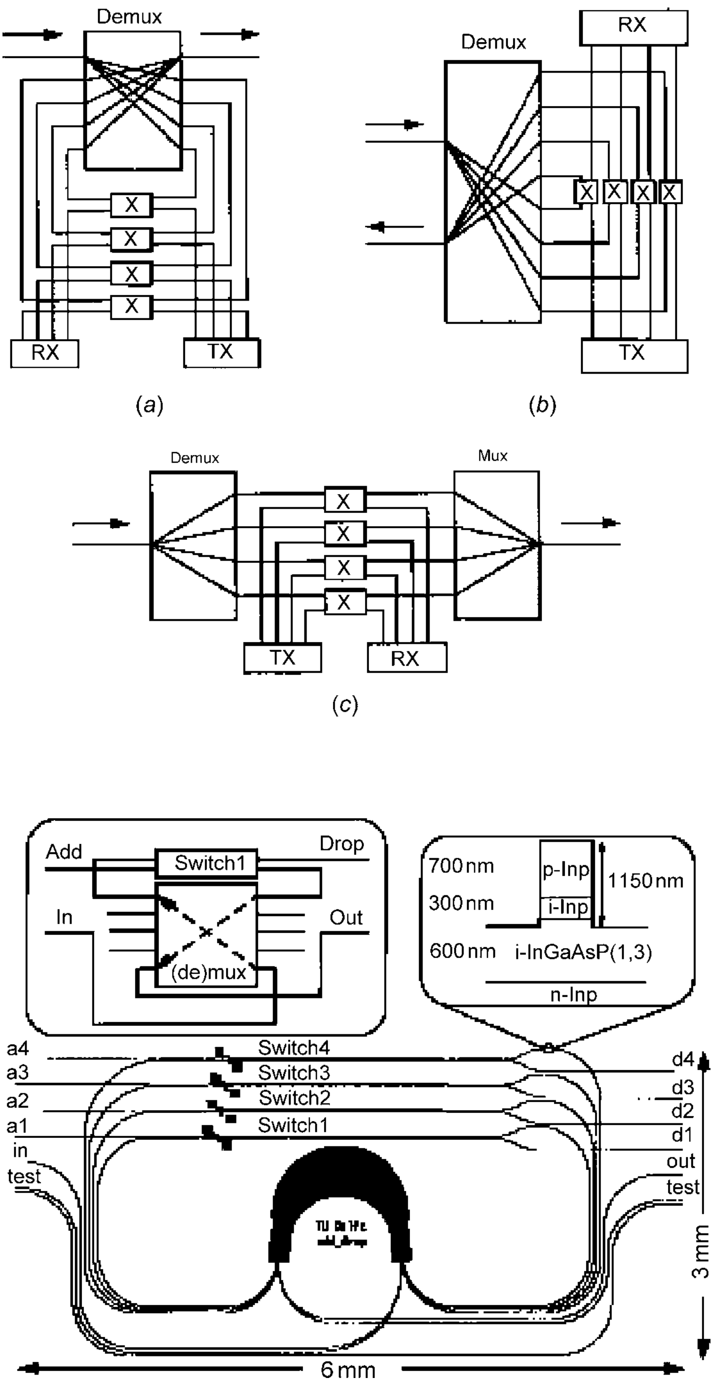

Demux
X
X
X
X
RX
TX
(a)
Demux
RX
X
X
X
X
TX
(b)
Demux
Mux
X
X
X
X
TX
RX
(c)
Add
Switch1
Drop
In
Out
(de)mux
700nm
300nm
p-Inp
i-Inp
1150nm
600nm
i-InGaAsP(1,3)
n-Inp
a4
a3
a2
a1
in
test
Switch4
Switch3
Switch2
Switch1
d4
d3
d2
d1
out
test
3mm
6mm

Figure B7.40. Schematic configurations for an add–drop multiplexer ADM using AWG for demultiplexing/multiplexing and space switches [17] (top): the conventional architecture (*c*) consists of two AWGs, and the fold back (*a*)/loop back (*b*) architectures use the same AWG for multiplexing and demultiplexing (bottom). Monolithic InP components for four channels with 400 GHz spacing [180]. Reproduced by permission of IEEE.

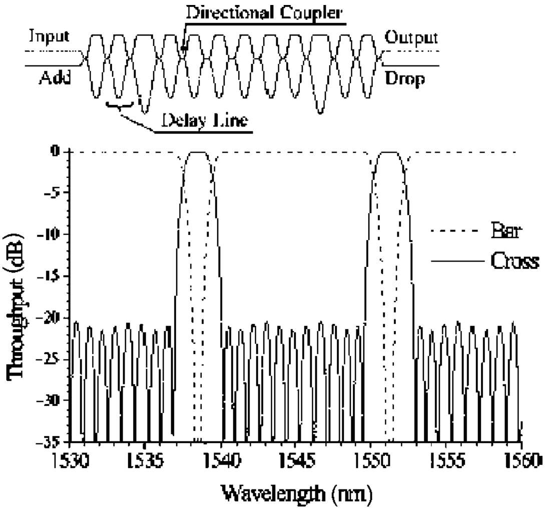


Figure B7.41. Serial lattice add–drop filter with 12 asymmetric Mach–Zehnder stages (top), theoretical device response for a drop (solid) and through channels (dashed) [44]. Reproduced by permission of IEEE.

in a positive or negative differential signal. A molecular absorption line serves as a wavelength reference, and a resolution of 10 MHz with sub-gigahertz accuracy was reported.

B7.5.7 All-optical signal processing—wavelength conversion and TDM switching

Like the all-optical switches described in section B7.5.4, all-optical signal processing [152] devices rely on third order nonlinearity $\chi(3)$. For practical all-optical devices, the nonlinearity has to be large enough to achieve a π phase shift with control pulses having picosecond length and picojoule energy, but at the same time the losses need to be sufficiently low. The semiconductor optical amplifier can overcome this trade-off, thus playing an important role in wavelength conversion, all-optical time division multiplexing and regeneration [186, 171].

For an SOA in the gain regime (i.e. with inverted population densities between valence and conduction bands), the following nonlinear phenomena occur.

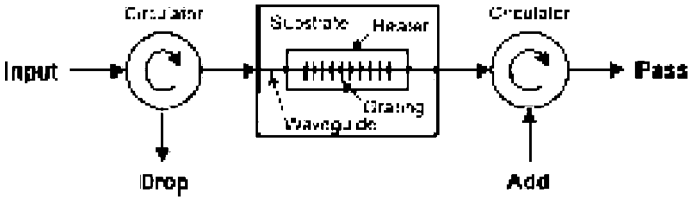


Figure B7.42. Optical add–drop filter using a thermo-optic tunable polymer Bragg grating and a pair of three port circulators [64]. The incoming signals will propagate through the left circulator into the Bragg grating, which reflects the drop wavelength back and passes the other channels through. In the same way, the add signals are reflected back and recombine with the through channels. Reproduced by permission of SPIE.

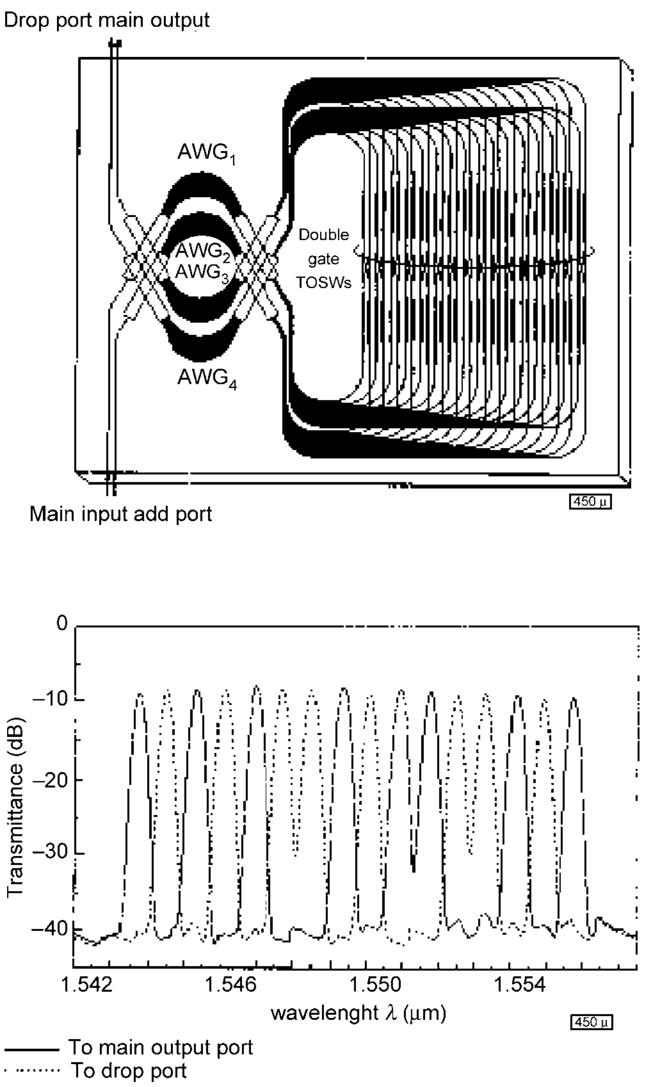


Figure B7.43. Integrated silica waveguide 2 × 2 optical crossconnect for 16 channels for each input/fibre [183]: layout with four AWGs and double stage TO switches, transmission spectrum for channels passing through (solid line) and dropped channels (dotted). Reproduced by permission of IEEE.

- Cross gain modulation (XGM)—an incoming photon will depopulate the conduction band, thus changing inversion and gain. One application of XGM is wavelength conversion, the concept of which is shown in figure B7.46: for an incoming '1' at the signal wavelength λ_{signal} the gain for the probe signal at wavelength λ_{probe} will drop, producing a '0'.
- Associated with the change in inversion is a variation in index and thus phase (cross phase modulation XPM), and this phase change is used in waveguide interferometers for signal processing:

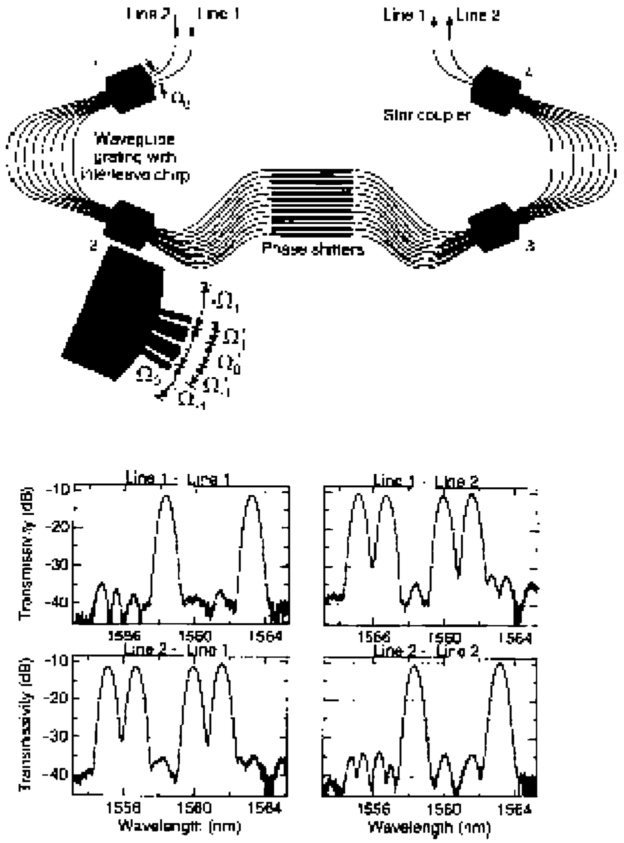
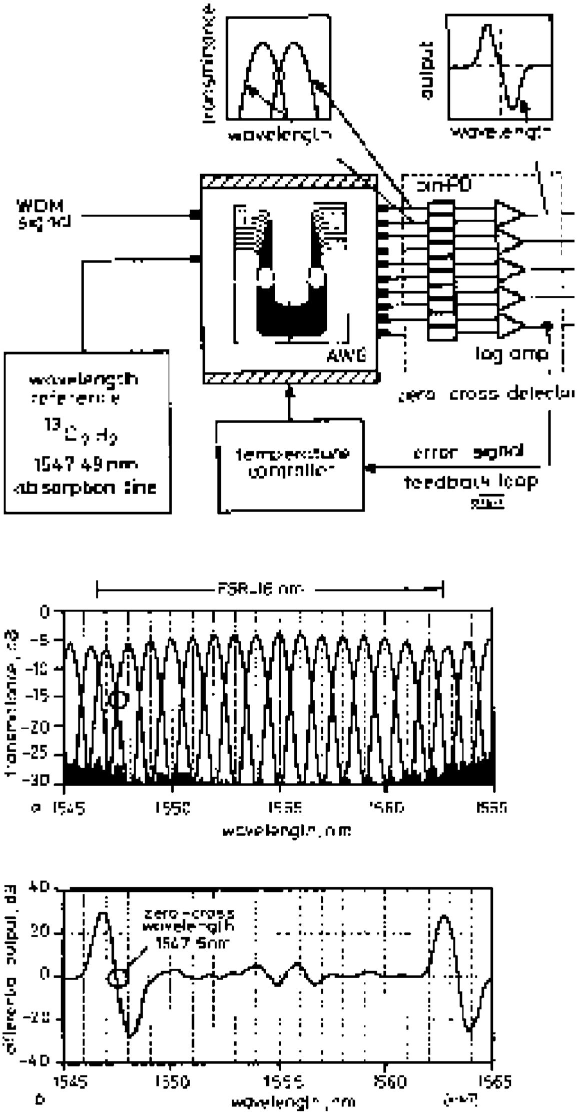

Line 2
Line 1
Line 1
Line 2
Waveguide grating with interleave chip
Star coupler
Phase shifters
1
2
3
4
Ω_1
Ω_1'
Ω_0
Ω_{-1}'
Ω_{-1}
Ω_2
Line 1 - Line 1
Line 1 - Line 2
Line 2 - Line 1
Line 2 - Line 2
Transmissivity (dB)
-10
-20
-30
-40
1556
1560
1564
Wavelength (nm)

Figure B7.44. Monolithic 2 × 2 crossconnect with two AWGs: design [184] and on-chip spectra of an InP device [177]. Reproduced by permission of IEEE.

- Time division demultiplexing: the monolithic InP MZI with SOAs in both arms depicted in figure B7.47 [187] (the architecture being the same as the GaAs all-optical passive switch in section B7.5.4/figure B7.39) represents a device for demultiplexing a 40 Gbit s^{-1} stream into 10 Gbit s^{-1} streams. Initially, the Mach–Zehnder is symmetric, but control pulse 1 will introduce a phase change in the upper SOA and will switch the signals from one output port to the other. As a result, the input data stream is demultiplexed. Control pulse 2 is introduced with an appropriate timing delay and resets the MZI to the initial state. The subsequent control pulses are needed for achieving switching speeds beyond the limit of the SOA carrier lifetimes (30–300 ps) [152].
- Wavelength conversion via XPM is explained schematically in figure B7.48(*a*): the switching signal $\lambda_{\mathrm{switch}}$ will introduce a de-phasing in the interferometer, thus impressing the data pattern on the continuous-wave probe wavelength $\lambda_{\mathrm{signal}}$. The result is one inverted and one noninverted signal as shown on the right hand side. The corresponding monolithic integrated InP device is shown in figure B7.48(*b*) together with the performance at 10 Gbit s^{-1} modulation [186], comparing the BER for the initial ('back-to-back') and wavelength converted signal. Besides monolithic solutions, hybrid integrated circuits had been published earlier [188].


transmittance
wavelength
output
wavelength
WDM signal
pinPD
AWG
log amp
zero cross detector
wavelength reference
$^{13}C_2H_2$
1547.49 nm absorption line
temperature controller
error signal
feedback loop
FSR=16 nm
transmittance, dB
0
-5
-10
-15
-20
-25
-30
a
1545
1550
1555
1560
1565
wavelength, nm
differential output, dB
40
20
0
-20
-40
zero-cross wavelength 1547.5nm
1545
1550
1555
1560
1565
b
wavelength, nm

a Transmittance of AWG
b Output of zero-cross detector

Figure B7.45. Wavelength monitoring circuit [185]: the large passbands of the AWG overlap produce an electrical differential signal for zero detection. Reproduced by permission of IEEE.

XPM and XGM wavelength conversion are both suitable for single wavelengths, but not in WDM systems as—besides the signal λ_{switch}—all the other data channels also contribute to XGM and XPM. Despite the fact that XPM devices have a more complicated layout and are difficult to fabricate, they have significant advantages: higher extinction ratio for the wavelength converted signal, availability of both inverted and noninverted data outputs and regeneration of the signal (see section B7.5.8). Also, the direct amplitude modulation in XGM introduces chirp

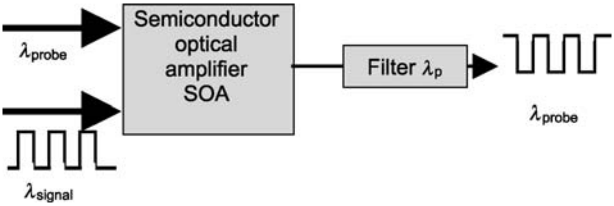


Figure B7.46. Wavelength conversion by cross-gain modulation (XGM) in a nonlinear semiconductor optical amplifier SOA. An incoming '1' on the signal wavelength will reduce the gain at the wavelength of the probe laser, producing an inverted signal.

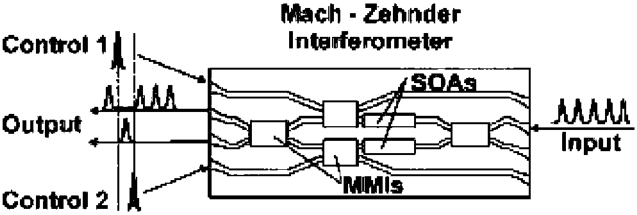


Figure B7.47. OTDM device [187]. The architecture is similar to the one in figure B7.39, but this time nonlinear InGaAsP SOAs in the 1.5 μm telecom window are used. Reproduced by permission of IEEE.

(i.e. the instantaneous phase of the signal is a function of the amplitude), but chirp can be controlled in XPM devices.

- Four wave mixing (FWM), which also arises in passive nonlinear transmission media such as fibre, causes three waves at frequencies ω_1, ω_2 and ω_3 to generate a fourth wave at frequency $\omega_1 + \omega_2 - \omega_3$. Pumping at wavelength $\omega_{pump} = \omega_1 = \omega_2$ and providing an idler at $\omega_{idler} = \omega_3$ results in a wavelength converted signal at $\omega_s = 2\omega_{pump} - \omega_{idler}$, as schematically shown in figure B7.49.

  In contrast to XPM and XGM, four wave mixing can convert several signals at once.

For an exhaustive treatment of all-optical time division multiplex system techniques, in particular using semiconductor and also fibre-based devices, see reference [63 chapter 9] and chapters A2.4, B.6 and C3.5.

For nonlinear optics, see chapters A2.4 and reference [8], and for SOAs in particular, see reference [7 chapter 11].

B7.5.8 All-optical regeneration

Degradation of the OSNR in a system arises from the accumulation of amplified spontaneous emission (ASE) from the line amplifiers, chromatic dispersion of the transmission fibre and components, PMD and third order nonlinearity of the fibre (self phase and cross phase modulation [189, 142]). This leads to a transparent reach of about 1500–2000 km, and the signal needs to be regenerated before further transmission. In [109], a re-circulating loop network with a reach of 69 000 km had been demonstrated using electronically driven modulators for reshaping the 20 Gbit s^{-1} signals for every 100 km.

All-optical regeneration could be an interesting alternative to electronic regeneration for different reasons: it is expected to be more cost effective through elimination of expensive O–E–O converters, moreover all-optical regeneration allows the processing of low as well as high bit rate signals rather than

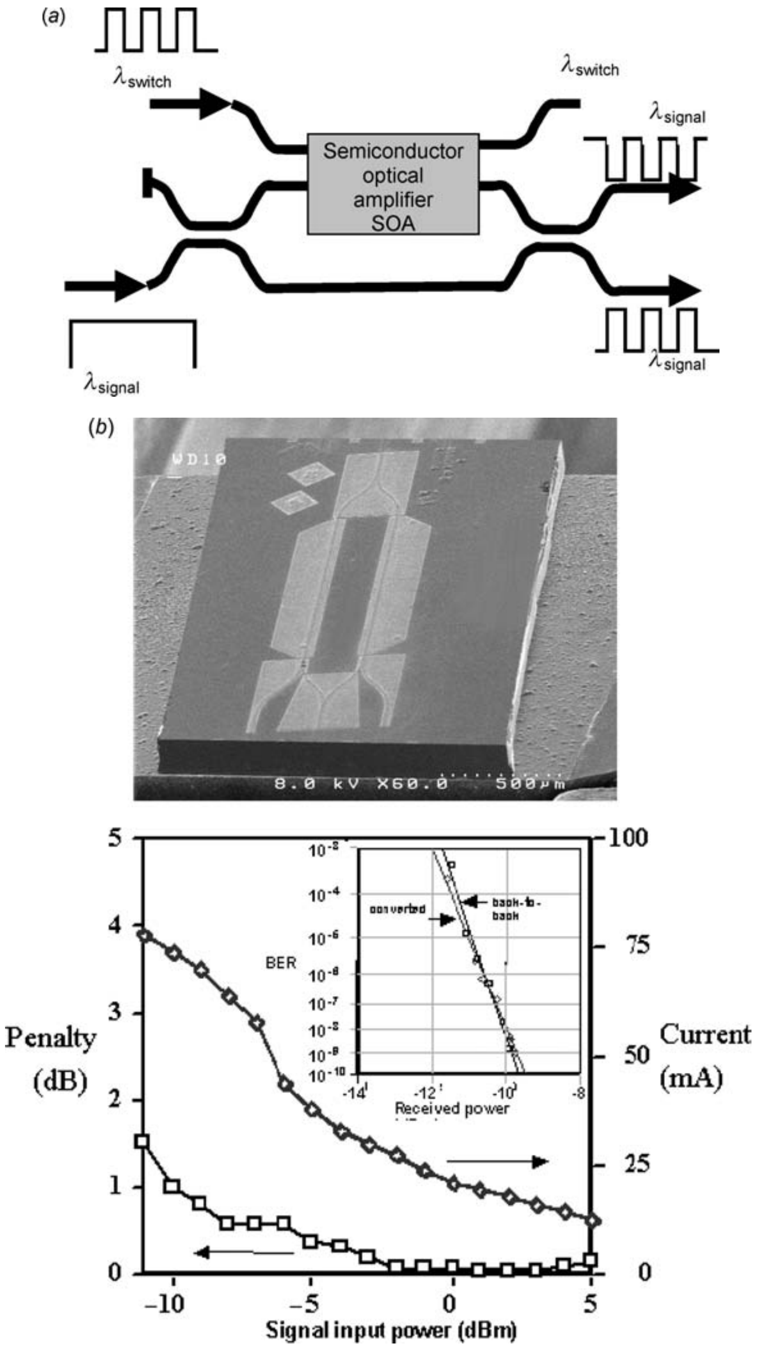

(a)
λ_{switch}
λ_{switch}
λ_{signal}
Semiconductor optical amplifier SOA
λ_{signal}
λ_{signal}
(b)
WD10
8.0 kV X60.0 500µm
Penalty (dB)
Current (mA)
Signal input power (dBm)
BER
converted
back-to-back
Received power
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
-14
-12
-10
-8
0
1
2
3
4
5
0
25
50
75
100
−10
−5
0
5

Figure B7.48. (*a*) Wavelength conversion through cross phase modulation XPM in nonlinear SOAs: the switching signal λ_{switch} will introduce a de-phasing in the interferometer, and the result is one inverted and one noninverted signal at wavelength λ_{signal}. (*b*) Monolithic InP cross phase modulation wavelength converter and performance in a 10 Gbit s^{-1} system; the inset compares the BER for the initial ('back-to-back') and wavelength converted signal [186]. Reproduced by permission of VDE Verlag.

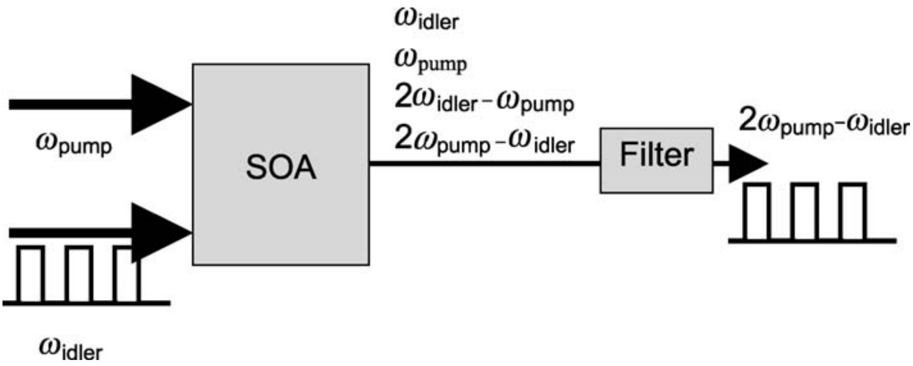

ω_{idler}
ω_{pump}
$2\omega_{idler}-\omega_{pump}$
$2\omega_{pump}-\omega_{idler}$
ω_{pump}
SOA
Filter
$2\omega_{pump}-\omega_{idler}$
ω_{idler}

Figure B7.49. Four-wave mixing in an SOA. The nonlinearity of the SOA will cause generation of sum and difference frequencies, and the signal is converted from frequency ω_i to $2\omega_{pump} - \omega_{idler}$.

being specifically designed for 10 or 40 Gbit s^{-1}. Also, 40 Gbit s^{-1} is believed to be the limitation for electronic regeneration but optical regeneration may perform at bit rates above this.

For regenerating a signal, three levels of manipulation are considered [107]:

- Optical re-amplification of the optical signal (1R regeneration).
- Amplification and re-shaping (2R).
- Amplification, re-shaping and retiming fully recovers the signal (3R). Retiming is necessary when accumulation of jitter, i.e. random signal delay through phase modulation in the fibre or in nonlinear devices (such as wavelength converters) is an issue.

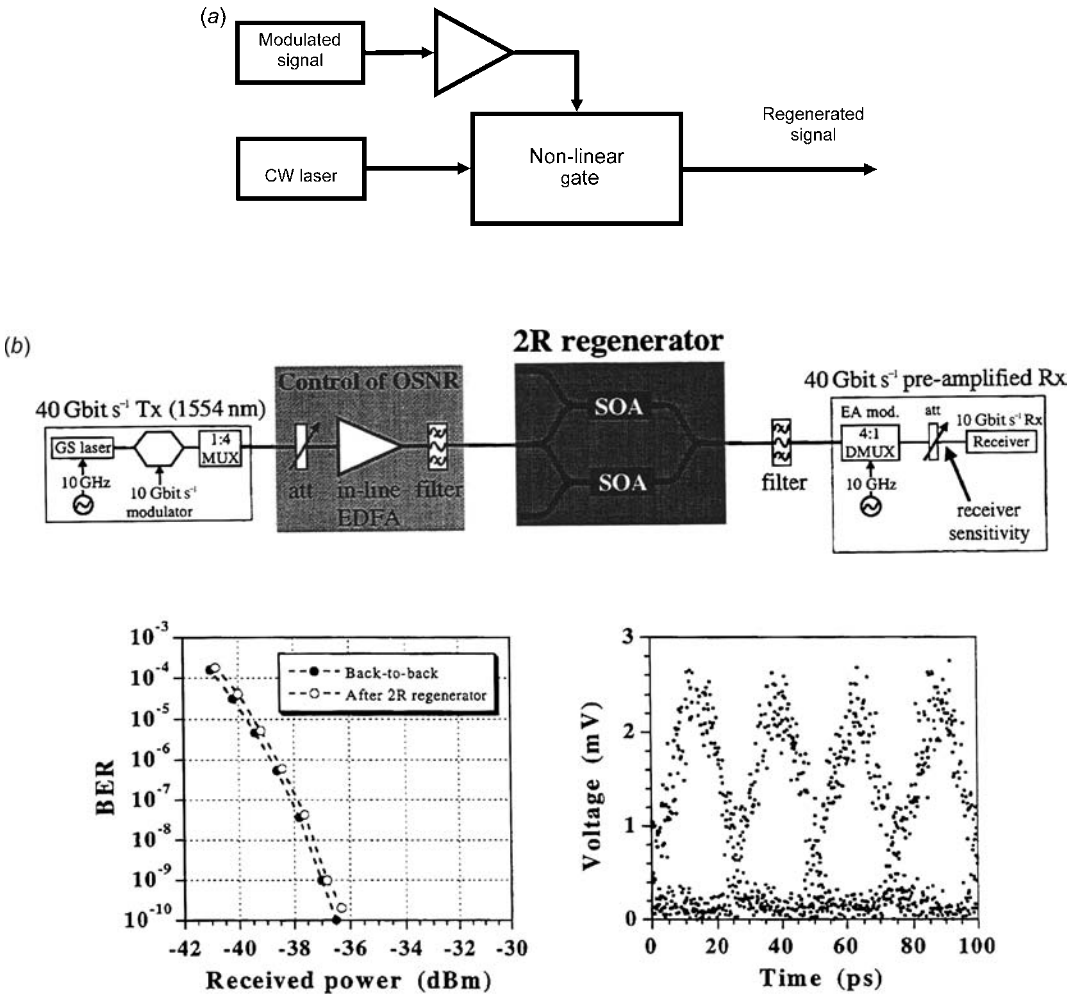
1R regeneration can be done all-optically in a linear amplifier (see chapters A1.6 and B5) or through OE-conversion followed by an electrical amplification and EO-conversion.

2R signal reshaping and amplification can be implemented with an amplifier and a nonlinear gate as sketched in figure B7.50(*a*): the nonlinear gate modulates a ‘clean’ CW signal having the same wavelength as the input data. Besides fibre based nonlinear devices, the SOA Mach–Zehnder interferometer previously used for XPM (section B7.5.7) provides a nonlinear amplitude response and can thus be used as a gate, as demonstrated in [190] for regeneration of 40 Gbit s^{-1} signals. In the SOA, the presence of the CW laser reduces the carrier lifetime through stimulated emission [152], thus enabling operation at high bit rates.

Figure B7.50(*b*) [190] shows the set-up for a 2R experiment at 40 Gbit s^{-1} as well as the BER results for a system with a 2R regenerator compared to the back-to-back operation of the Tx/Rx. The penalty is <0.2 dB, indicating that the 2R regenerator restores the signal almost to its initial quality.

3R regeneration (figure B7.51(*a*)) requires—besides 2R amplification and reshaping—the extraction of a jitter-free clock signal from the incoming data stream. The clock signals are then sent to the nonlinear optical gate, which is modulated by the initial data stream, and as a result a jitter-free signal is generated at the 3R output.

The main difficulty is the extraction of the clock signal, which can be done opto-electronically or all-optically via, for example, self-pulsating laser diodes (SP-LDs) [191, 192]. The SP-LD initially has a free-running repetition rate, and the basic principle (figure B7.51(*b*)) of clock extraction is that the SP-LD will change and lock its repetition rate to the data bit rate when optical data signals are injected into the



(a) (b)

Figure B7.50. (*a*) 2R regenerator. The nonlinear all-optical gate modulates a 'clean' CW signal having the same wavelength as the modulated input signal. (*b*) Set-up for 2*R* regeneration at 40 Gbit s^{-1} (top) [190] and system performance (bottom): bit error rate versus received power in a back-to-back configuration and after the 2R regenerator. The penalty of <0.2 dB indicates that the signals are regenerated almost with their initial signal-to-noise ratio. The bottom right figure is the oscilloscope trace of the received signal. Reproduced by permission of the Optical Society of America.

laser. The oscilloscope traces in figure B7.51(*c*) [193] show the incoming 5 Gbit s^{-1} data consisting of 1s and 0s (top) and the extracted clock signal constists of a train of '1' pulses.

In system experiments, 3R regeneration has been demonstrated for up to 40 Gbit s^{-1} bit rate using polarization rotation in a nonlinear fibre [194] and later on using a fibre based interferometer including an SOA as the nonlinear decision gate [195]. The latter set-up and experimental results are shown in figure B7.51(*d*): the inserts represent oscilloscope traces of the incoming and regenerated signals, and the BER diagrams show a PP of about 2.2 dB.

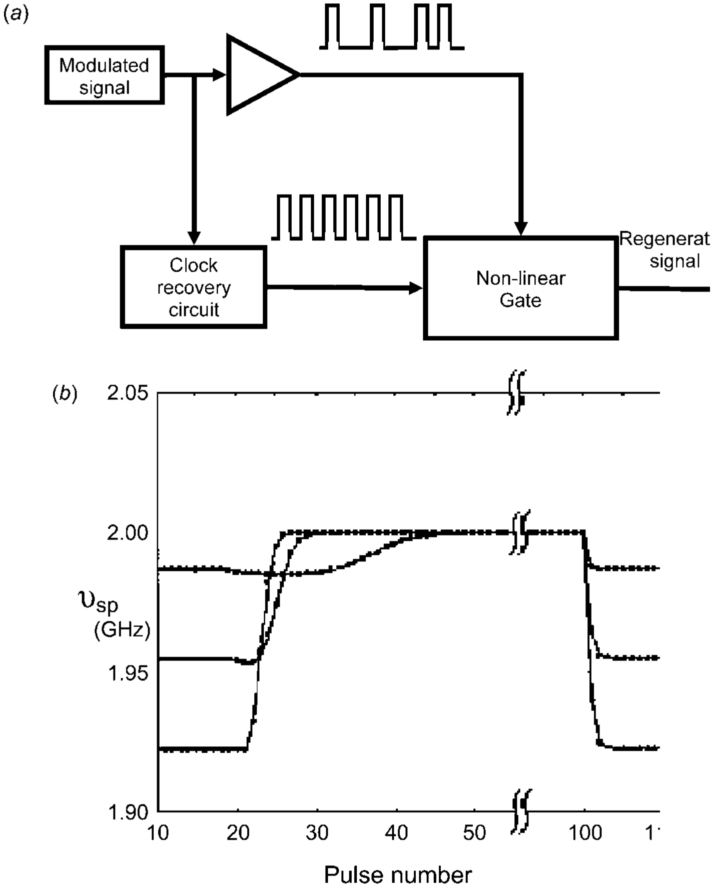


Figure B7.51. (*a*) Schematic 3R regenerator, consisting of a (noisy) input signal, from which the clock recovery circuit will extract a regular train of signals. The clock signals in turn will gate the all-optical nonlinear circuit, so that the data are not only re-amplified and re-shaped but also re-timed. (*b*) Self-pulsating laser diode [192] for clock recovery: the figure shows how laser diodes with different free-running frequencies will lock their repetition rate when synchronization pulses of precisely 2 GHz are injected. (*c*) Self-pulsating laser diode as a 5 Gbit s^{-1} clock recovery circuit: the clock signal (a train of '1' pulses) is extracted from an incoming signal consisting of '1's and '0's (top) [193]. (*d*) 3R regeneration (40 Gbit s^{-1}) [195]. In the experimental set-up (top), the nonlinear gate consists of a fiber-based Mach–Zehnder interferometer with a nonlinear SOA in one arm. BER (bottom) as a function of the received power, comparing the regenerated signals (open symbols) to a back-to-back experiment (full symbols). For a BER of 10^{-9}, the power penalty is about 2.2 dB. The insets are oscilloscope traces of the initial and regenerated signals. Reproduced by permission of IEEE.

B7.6 Conclusion

Integrated optics deals with compact single function devices, devices with integrated multiple functions (either on a single chip or in a single package) or devices with a functionality which cannot be achieved with common bulk or thin film optic devices. Besides these advantages in functionality or performance, the main driver for integrated optic devices is the desire for device footprint and reduced cost. The latter will be achieved through ongoing investments and standardization.

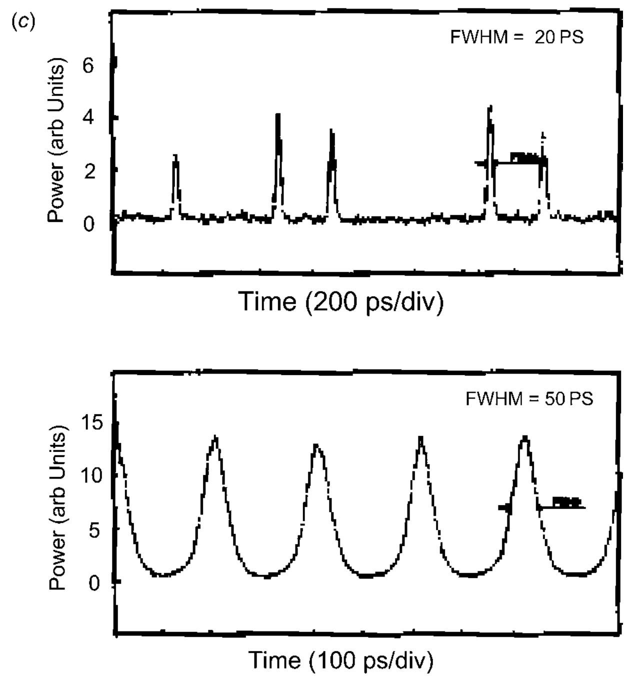


Figure B7.51. (Continued)

Tremendous work has been done over the last 20 years, but there are still a few words of caution:

- The technology is much less mature than electronics.
- As pointed out in section B7.3, there are multiple materials and technologies: that unlike in electronics where Si has an outstanding role, at present there is no material that is capable of addressing all needs in terms of device performance and cost simultaneously. InP may take the lead as it is suitable for the fabrication of both active and passive functions, but fundamental issues such as wafer size and materials compatibility will need to be addressed.
- Another difference with respect to electronics relates to the density of the optical circuits: the typical size of an optical chip is a few mm^2, which considerably limits the number of functions on a single substrate. The connection of the optical circuit to the outside world is done by aligning single mode fibre—the typical fibre diameter of 125 μm limits the number of I/O connections.
- Many current high speed devices (such as the switches in section B7.5.4, or the optical crossconnect in B7.5.5) use the electro-optic effect in SOAs, but all-optical devices (in analogy to the 'all-electronic' transistor) are still a subject of fundamental research, and thus are far from deployment.

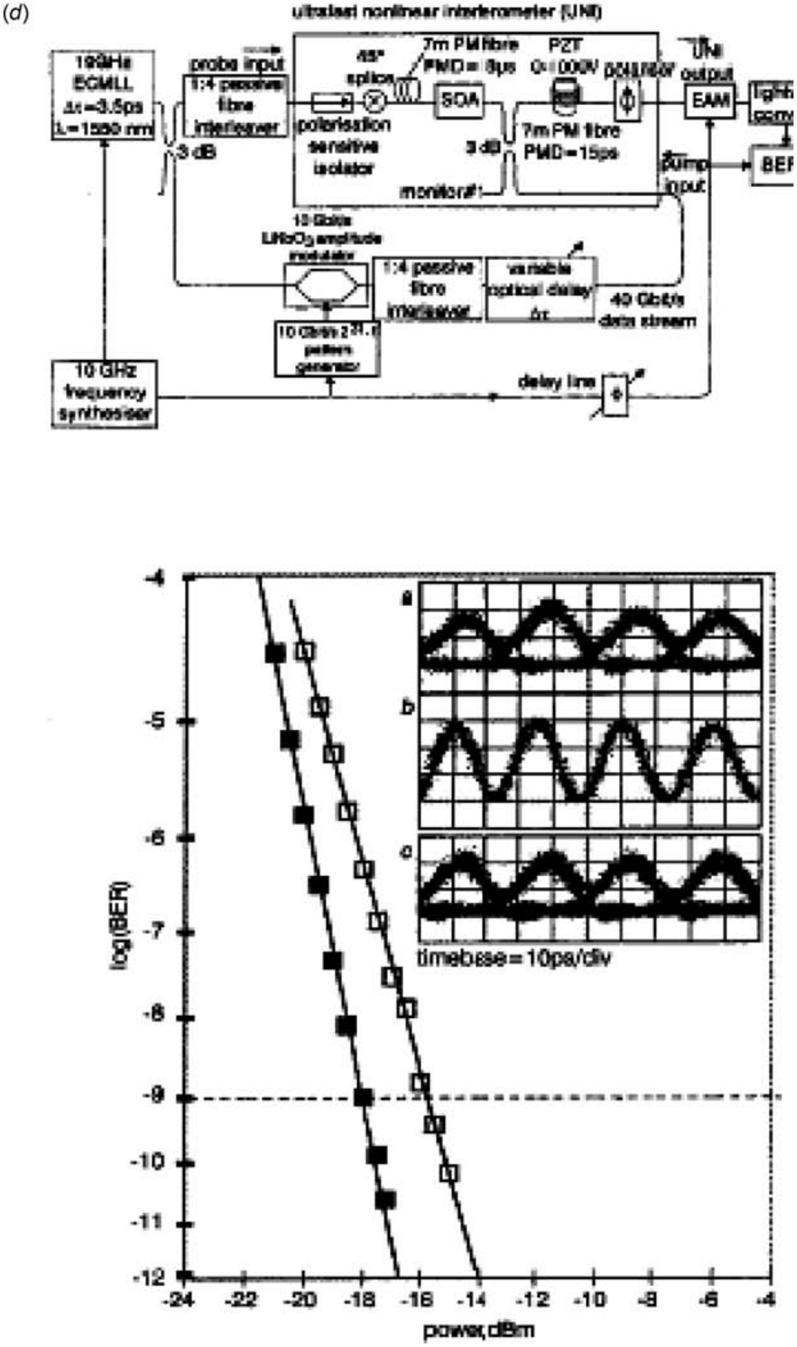

(d)
ultrafast nonlinear interferometer (UNI)
10GHz ECMLL Δτ=3.5ps λ=1550 nm
3 dB
probe input
1:4 passive fibre interleaver
polarisation sensitive isolator
45° splice
7m PM fibre PMD=8ps
SOA
3 dB
monitor#1
PZT 0-1000V
7m PM fibre PMD=15ps
polariser
UNI output
EAM
pump input
BER
10 Gbit/s LiNbO3 amplitude modulator
1:4 passive fibre interleaver
variable optical delay Δτ
40 Gbit/s data stream
10 Gbit/s 2^31-1 pattern generator
10 GHz frequency synthesiser
delay line
log(BER)
-4
-5
-6
-7
-8
-9
-10
-11
-12
power, dBm
-24
-22
-20
-18
-16
-14
-12
-10
-8
-6
-4
a
b
c
timebase=10ps/div

Figure B7.51. (Continued)

However, there are key advantages of optical devices: the capability of achieving modulation speeds beyond 40 Gbit s^{-1}, the absence of electromagnetic interference, and finally optical devices may be used for sensing applications in environments in which electrical charges are prohibited.

Today, integrated optics is the technology of choice for a few functions such as filters, switches and gain equalizers. In future telecommunication applications, we will see an increasing deployment due to the advantages in device functionality, footprint, device cost and finally modulation speed resulting in higher transmission capacities.

References

[1] Ladouceur F and Love J D 1996 *Silica-Based Buried Channel Waveguides and Devices* (London: Chapman and Hall)

[2] Okamoto K 2000 *Fundamentals of Optical Waveguides* 2nd edn English Translation (San Diego: Academic)

[3] Kawachi M 1990 Silica waveguides on silicon and their application to integrated-optic components *Opt. Quantum Electron.* **22** 391

[4] Eldada L and Shacklette L W 2000 Advances in polymer integrated optics *IEEE J. Sel. Top. Quantum Electron.* **6** 54

[5] Cocorullo G and Rendina I 1992 Thermo-optical modulation at 1.5 μm in silicon *Electron. Lett.* **28** 83

[6] Jackson J D 1998 *Classical Electrodynamics* 3rd edn (New York: Wiley)

[7] Murphy E J (ed) 1999 *Integrated Optical Circuits and Components Design and Applications* (New York: Dekker)

[8] Sutherland R L 1996 *Handbook of Nonlinear Optics* (New York: Dekker)

[9] Alferness R 1988 *Titanium-Diffused Lithium Niobate Waveguide Devices Guided Wave Optoelectron.*, ed T Tamir (New York: Springer)

[10] Ikushima A J, Fujiwara T and Saito K 2000 Silica glass: a material for photonics *J. Appl. Phys.* **88** 1201

[11] Wang W, Shi Y, Olson D J, Lin W and Bechtel J H 1997 Polymer integrated modulators for photonic data link applications *Proc. SPIE—The International Soc. for Opt. Eng.* vol 2997 p 114

[12] Takagi A, Jinguji K and Kawachi M 1992 Silica-based waveguide-type wavelength-insensitive couplers (WINCs) with series-tapered coupling structure *J. Lightwave Technol.* **10** 1814

[13] Nishihara H, Haruna M and Suhara T 1989 *Optical Integrated Circuits* 2nd edn (New York: McGraw-Hill)

[14] Dragone C 1989 Efficient $N \times N$ star couplers using Fourier optics *IEEE J. Lightwave Technol.* **7** 479

[15] Okamoto K, Okazaki H, Ohmori Y and Kato K 1992 Fabrication of large scale integrated-optic $N*N$ star couplers *IEEE Photon. Technol. Lett.* **4** 1032

[16] Soldano L B and Pennings E C M 1995 Optical multi-mode interference devices based on self-imaging: principles and applications *J. Lightwave Technol.* **13** 615

[17] Smit M K and Van Dam C 1996 PHASAR-based WDM-devices: principles design and applications *IEEE J. Sel. Top. Quantum Electron.* **2** 236

[18] Takahashi H, Oda K, Toba H and Inoue Y 1995 Transmission characteristics of arrayed waveguide $N*N$ wavelength multiplexer *J. Lightwave Technol.* **13** 447

[19] Soref R A and Little B E 1998 Proposed N-wavelength M-fibre WDM crossconnect switch using active microring resonators *IEEE Photon. Technol. Lett.* **10** 1121

[20] März R 1995 *Integrated Optics Design and Modeling* (Boston: Artech)

[21] Lenz G, Eggleton B J and Madsen C K 1999 Optical filter dispersion in WDM systems: a review *OSA Trends in Optics and Photonics WDM Components* vol 29 p 246

[22] Madsen C K and Zhao J H 1999 *Optical Filter Design and Analysis A Signal Processing Approach* (New York: Wiley)

[23] Jinguji K 1998 Broadband programmable optical frequency filter *Electron. Commun. Jpn Part 2 (Electronics)* **81** 1

[24] Madsen C K and Zhao J H 1996 A general planar waveguide autoregressive optical filter *J. Lightwave Technol.* **14** 437

[25] Madsen C K 2000 General IIR optical filter design for WDM applications using all-pass filters *J. Lightwave Technol.* **18** 860

[26] Himeno A, Kato K and Miya T 1998 Silica-based planar lightwave circuits *IEEE J. Sel. Top. Quantum Electron.* **4** 913

[27] Martinu L and Poitras D 2000 Plasma deposition of optical films and coatings: a review *J. Vacuum Sci. Technol. A* **18** 2619

[28] Bushan B, Murarka S P and Gerlach J 1990 Stress in silicon dioxide films deposited using chemical vapour deposition techniques and the effect of annealing on these stresses *J. Vacuum Sci. Technol. B* **8** 1068

[29] Kilian A, Kirchof J, Przyrembel G and Wischmann W 2000 Birefringence free planar optical waveguide made by flame hydrolysis deposition (FHD) through tailoring of the overcladding *J. Lightwave Technol.* **18** 193

[30] Ojha S M, Cureton C, Bricheno T, Day S, Moule D, Bell A J and Taylor J 1998 Simple method of fabricating polarisation-insensitive and very low crosstalk AWG grating devices *Electron. Lett.* **34** 78

[31] Coburn J W and Winters H F 1979 Plasma etching—a discussion of mechanisms *J. Vacuum Sci. Technol.* **16** 391

[32] Vossen J-L and Kern W 1978 *Thin Film Processes* (Orlando: Academic)

[33] Ramaswamy R U and Srivastava R 1988 Ion-exchanged glass waveguides: a review *J. Lightwave Technol.* **6** 984

[34] Cheng H C and Ramaswamy R U 1990 A dual wavelength directional coupler demultiplexer by ion exchange in glass *IEEE Photon. Technol. Lett.* **2** 637

[35] Mizrahi V, Lemaire P J, Erdogan T, Reed W A, DiGiovanni D J and Atkins R M 1993 Ultraviolet laser fabrication of ultrastrong optical fibre gratings and of germania-doped channel waveguides *Appl. Phys. Lett.* **63** 1727

[36] Herman P R, Chen K, Li J, Wie M, Ihlemann J and Marowsky G 2001 F_2 lasers: precise shaping and trimming of photonic components *2001 Digest of LEOS Summer Topical Meetings: Advanced Semiconductor Lasers and Applications/Ultraviolet and Blue Lasers and their Applications/Ultralong Haul DWDM Transmission and Networking/WDM Components*

[37] Strasser T A, Erdogan T, White A E, Mizrahi V and Lemaire P J 1994 Ultraviolet laser fabrication of strong nearly polarization-independent Bragg reflectors in germanium-doped silica waveguides on silica substrates *Appl. Phys. Lett.* **65** 3308

[38] Bazylenko M V, Gross M, Chu P L and Moss D 1996 Photosensitivity of Ge-doped silica deposited by hollow cathode PECVD *Electron. Lett.* **32** 1198

[39] Bazylenko M V, Gross M and Moss D 1997 Mechanisms of photosensitivity in germanosilica films *J. Appl. Phys.* **81** 7497

[40] Maxwell G D, Ainslie B J, Williams D L and Kashyap R 1993 UV written 13 dB reflection filters in hydrogenated low loss planar silica waveguides *Electron. Lett.* **29** 425

[41] Maxwell G D and Ainslie B J 1995 Demonstration of a directly written directional coupler using UV-induced photosensitivity in a planar silica waveguide *Electron. Lett.* **31** 95

[42] Hoffmann M, Kopka P and Voges E 1997 Low-loss fibre-matched low-temperature PECVD waveguides with small-core dimensions for optical communication systems *IEEE Photon. Technol. Lett.* **9** 1238

[43] de Ridder R M, Warhoff K, Driessen A, Lambeck P V and Albers H 1998 Silicon oxynitride planar waveguiding structures for application in optical communication *IEEE J. Sel. Top. Quantum Electron.* **4** 930

[44] Offrein B J, Horst F, Bona G L, Salemink H W M, Germann R and Beyeler R 1999 Wavelength tunable 1-from-16 and flat passband 1-from-8 add–drop filters *IEEE Photon. Technol. Lett.* **11** 1440

[45] Mertens H, Andersen K N and Svendsen W E 2002 Optical loss analysis of silicon rich nitride waveguides *ECOC'02: 28th European Conf. on Optical Communication* poster 1.38

[46] Jalali B, Yegnanarayanan S, Yoon T, Yoshimoto T, Rendina I and Coppinger F 1998 Advances in silicon-on-insulator optoelectronics *IEEE J. Sel. Top. Quantum Electron.* **4** 93

[47] Rickman A G and Reed G T 1994 Silicon-on-insulator optical rib waveguides: loss mode characteristics bends and y-junctions *IEE Proc. Optoelectron.* **141** 391

[48] Tang C K, Kewell A K, Reed G T, Rickman A G and Namavar F 1996 Development of a library of low-loss silicon-on-insulator optoelectronic devices *IEE Proc. Optoelectron.* **143** 312

[49] Cocorullo G, Della Corte F G, de Rosa R, Rendina I, Rubino A and Terzini E 1998 Amorphous silicon-based guided-wave passive and active devices for silicon integrated optoelectronics *IEEE J. Sel. Top. Quantum Electron.* **4** 997

[50] Soref R A, Schmidtchen J and Petermann K 1991 Large single-mode rib waveguides in GeSi–Si and Si-on-SiO_2 *IEEE J. Quantum Electron.* **27** 1971

[51] Bestwick T 1998 ASOC™-a silicon-based integrated optical manufacturing technology *Proc. 48th Electronic Components and Technol. Conf. (Cat No 98CH36206)* p 566

[52] Fischer U, Zinke T, Kropp J-R, Arndt F and Petermann K 1996 0.1 dB/cm waveguide losses in single-mode SOI rib waveguides *IEEE Photon. Technol. Lett.* **8** 647

[53] Joannopoulos J D, Villeneuve P R and Shanhui Fan 1997 Photonic crystals: putting a new twist on light *Nature* **386** 143

[54] Berger V 1999 From photonic band gaps to refractive index engineering *Opt. Mater.* **11** 131

[55] Pirnat T and Friedman L 1991 Electro-optic mode-displacement silicon light modulator *J. Appl. Phys.* **70** 3355

[56] Soref R A and Bennett B R 1987 Electrooptical effects in silicon *IEEE J. Quantum Electron.* **QE-23** 123

[57] Jackson S M, Hewitt P D, Reed G T, Tang C K, Evans A G R, Clark J, Aveyard C and Namavar F 1998 A novel optical phase modulator design suitable for phased arrays *J. Lightwave Technol.* **16** 2016

[58] Wooten E L *et al* 2000 A review of lithium niobate modulators for fibre-optic communications systems *IEEE J. Sel. Top. Quantum Electron.* **6** 69

[59] Kip D 1998 Photorefractive waveguides in oxide crystals: fabrication properties and applications *Appl. Phys. B (Lasers and Opt.)* **B67** 131

[60] Korkishko Y N and Federov V A 1999 *Ion Exchange in Single Crystals for Integrated Optics and Optoelectronics* (Cambridge: Cambridge International Science Publishing)

[61] Hofmann D, Schreiber G, Haase C, Herrmann H, Grundkotter W, Ricken R and Sohler W 1999 Quasi-phase-matched difference-frequency generation in periodically poled Ti:$LiNbO_3$ channel waveguides *Opt. Lett.* **24** 896

[62] Noguchi K, Mitomi O and Miyazawa H 1996 Low-voltage and broadband Ti:$LiNbO_3$ modulators operating in the millimeter wavelength region *OFC'96: Optical Fiber Communication 1996 Technical Digest Series Conf. Ed. (IEEE Cat No96CH35901)* vol 2 p 205

[63] Grote N and Venghaus H (ed) 2001 *Fibre Optic Communication Devices* (Heidelberg: Springer)

[64] Eldada L, Blomquist R, Shacklette L W and McFarland M J 2000 High-performance polymeric componentry for telecom and datacom applications *Opt. Eng.* **39** 596

[65] Bauer H-D, Ehrfeld W, Harder M, Paatzsch T, Popp M and Smaglinski I 2000 Polymer waveguide devices with passive pigtailing: an application of LIGA technology *Synth. Met.* **115** 13

[66] Viens J-F, Callender C L, Noad J P and Eldada L 2000 Compact wide-band polymer wavelength-division multiplexers *IEEE Photon. Technol. Lett.* **12** 1010

[67] Keil N *et al* 2000 Thermo-optic vertical coupler switches using hybrid polymer/silica integration technology *Electron. Lett.* **36** 430

Keil N *et al* 2000 Thermo-optic switches using vertically coupled polymer/silica waveguides *Proc. 26th European Conference on Opt. Commun.* p 101

[68] Inoue Y, Kaneko A, Hanawa F, Takahashi H, Hattori K and Sumida S 1997 Athermal silica-based arrayed-waveguide grating multiplexer *Electron. Lett.* **33** 1945

[69] Kokubun Y, Takizawa M and Taga S 1994 Three-dimensional athermal waveguides for temperature independent lightwave devices *Electron. Lett.* **30** 1223

[70] Moroni M and Vallon S 1999 Athermalized polymer overclad integrated planar optical waveguide device and its manufacturing method EP 1026526 A1

[71] Dalton L, Harper A, Ren A, Wang F, Todorova G, Chen J, Zhang C and Lee M 1999 Polymeric electro-optic modulators: from chromophore design to integration with semiconductor very large scale integration electronics and silica fibre optics *Ind. Eng. Chem. Res.* **38** 8

[72] Wiesmann R, Kalveram S, Rudolph S, Johnck M and Neyer A 1996 Singlemode polymer waveguides for optical backplanes *Electron. Lett.* **32** 2329

[73] Kik P G and Polman A 1998 Erbium-doped optical-waveguide amplifiers on silicon *MRS Bull.* **23** 48

[74] Shooshtari A, Meshkinfam P, Touam T, Andrews M and Najafi S 1998 Ion-exchanged Er/Yb phosphate glass waveguide amplifiers and lasers *Opt. Eng.* **37** 1188

[75] Poulsen M 2002 Private communication

[76] Bonar J, Bebbington J A, Aitchison J S, Maxwell G D and Ainslie B J 1995 Aerosol doped Nd planar silica waveguide laser *Electron. Lett.* **31** 99

[77] Baumann I, Bosso S, Brinkmann R, Corsini R, Dinand M, Greiner A, Schafer K, Sochtig J, Sohler W, Suche H and Wessel R 1996 Er-doped integrated optical devices in $LiNbO_3$ *IEEE J. Sel. Top. Quantum Electron.* **2** 355

[78] Polman A 1997 Erbium implanted thin film photonic materials *J. Appl. Phys.* **82** 1

[79] Mickelson A R, Basavanhally N R and Cheng Lee Yung (ed) 1997 *Optoelectronic Packaging* (New York: Wiley)

[80] Hunsperger R G 1995 *Integrated Optics Theory and Technology* 4th edn (Heidelberg: Springer)

[81] Kaufmann H, Buchmann P, Hirter R, Melchior H and Guekos G 1986 Self-adjusted permanent attachment of fibres to GaAs waveguide components *Electron. Lett.* **22** 642

[82] GR-20-CORE *Generic Requirements for Optical Fiber and Optical Fiber Cable* http://www.telcordia.com

[83] Paris B 2000 Multiple planar complex optical devices and the process of manufacturing the same EP 1170606 A1

[84] Shimizu N, Imoto N and Ikeda M 1983 Fusion splicing between optical circuits and optical fibres *Electron. Lett.* **19** 96

[85] Modavis R and Webb T 1994 Anamorphic microlens for coupling optical fibres to elliptical light beams US 5455879 A1

[86] Kobayashi M and Kato K 1994 Hybrid optical integration technology *Electron. Commun. Jn Part 2 (Electron.)* **77** 67

[87] Lealman I F, Kelly A E, Rivers L J, Perrin S D and Moore R 1998 Improved gain block for long wavelength (1,55mum) hybrid integrated devices *Electron. Lett.* **34** 2247

[88] Büttgenbach S 1993 *Mikromechanik* (Stuttgart: Teubner Studienbücher)

[89] Krishnamoorthy A V and Goossen K W 1998 Optoelectronic-VLSI: photonics integrated with VLSI circuits *IEEE J. Sel. Top. Quantum Electron.* **4** 899

[90] Wale M and Goodwin M 1992 Flip-chip bonding optimizes opto-ICs *IEEE Circuits Devices Mag.* **8** 25

[91] GR-468-CORE *Generic Reliability Assurance Requirements for Optoelectronic and Electro-Opto-Mechanical Devices Used in Telecommunications* http://www.telcordia.com

[92] Mino S, Yoshino K, Yamada Y, Terui T, Yasu M and Moriwaki K 1995 Planar lightwave circuit platform with coplanar waveguide for opto-electronic hybrid integration *J. Lightwave Technol.* **13** 2320

[93] Yoshida J 1999 Hybridization of active and passive optical devices toward multifunctional optical modules *ECOC'99: Proceedings 25th European Conf. on Opt. Communication* p 170

[94] Terui H, Shimokozono M, Yanagisawa M, Hashimoto T, Yamada Y and Horiguchi M 1996 Hybrid integration of eight channel PD-array on silica-based PLC using micro-mirror fabrication technique *Electron. Lett.* **32** 1662

[95] Lai Q, Hunziker W and Melchior H 1996 Silica on Si waveguides for self-aligned fibre array coupling using flip-chip Si V-groove technique *Electron. Lett.* **32** 1916

[96] Collins J V, Lealman I F, Kelly A and Ford C W 1997 Passive alignment of second generation optoelectronic devices *IEEE J. Sel. Top. Quantum Electron.* **3** 1441

[97] Choi M H, Koh H J, Yoon E S, Shin K C and Song K C 1999 Self-aligning silicon groove technology platform for the low cost optical module *1999 Proc. Conf. on 49th Electronic Components and Technology* p 1140

[98] Qing Tan and Lee Y C 1996 Soldering technology for optoelectronic packaging *1996 Proc. 46th Electron. Components and Technol. Conf.* p 26

[99] Rinne G A 1997 Solder bumping methods for flip chip packaging *1997 Proc. 47th Electron. Components and Technol. Conf.* p 240

[100] Jackson K P, Flint E B, Cina M F, Lacey D, Trewhella J M, Caulfield T and Sibley S 1992 A compact multichannel transceiver module using planar-processed optical waveguides and flip-chip optoelectronic components *1992 Proc. 42nd Electronic Components and Technol. Conf.* p 93

[101] Sang-Hwan Lee, Gwan-Chong Joe, Ki-Sung Park, Hong-Man Kim, Dong-Goo Kim and Hyung-Moo Park 1995 Optical device module packages for subscriber incorporating passive alignment techniques *1995 Proc. 45th Electron. Components and Technol. Conf.* p 841

[102] Basavanhally N 1993 Application of soldering technologies for opto-electronic component assembly *Advances in Electronic Packaging 1993 Proc. 1993 ASME Int. Electron. Packaging Conf.* p 1149

[103] Lee C C, Wang C Y and Matijasevic G 1993 Advances in bonding technology for electronic packaging *Trans. ASME J. Electron. Packaging* **115** 201

[104] Bigo S *et al* 2001 10.2 Tbit/s (256 × 427 Gbit/s PDM/WDM) transmission over 100 km TeraLight™ fibre with 128 bit/s/Hz spectral efficiency *OFC 2001. Optical Fiber Communication Conference and Exhibit Technical Digest Postconference Edition Postdeadline Papers* p PD 25

[105] Desurvire E 1994 *Erbium-Doped Fiber Amplifiers: Principles and Applications* 1st edn (New York: Wiley–Interscience)

[106] Agrawal G 1995 *Nonlinear Fiber Optics* 2nd edn (San Diego: Academic)

[107] Simon J C, Billess L, Dupas A and Bramerie L 1999 All optical regeneration techniques *ECOC'99 25th European Conf. on Optical Commun.* p 256

[108] Pender W A, Watkinson P J, Greer E J and Ellis A D 1995 10 Gbit/s all-optical regenerator *Electron. Lett.* **31** 1587

[109] Ellis A D and Widdowson T 1995 690 node global OTDM network demonstration *Electron. Lett.* **31** 1171

[110] Kuznetsov M, Froberg N M, Henion S R and Rauschenbach K A 1999 Power penalty for optical signals due to dispersion slope in WDM filter cascades *IEEE Photon. Technol. Lett.* **11** 1411

[111] Hida Y, Jinguji K and Takato N 1998 Wavelength demulti/multiplexers with non-sinusoidal filtering characteristics composed of point-symmetrically connected Mach–Zehnder interferometers *Electron. Commun. Jn Part 2 (Electronics)* **81** 19

[112] Chiba T, Arai H, Ohira K, Nonen H, Okano H and Uetsuka H 2001 Novel architecture of wavelength interleaving filter with Fourier transform-based MZIs *OFC 2001: Optical Fiber Communication Conference and Exhibit Technical Digest Postconference Edition* p WB5

[113] Oguma M, Jinguji K, Kitoh T, Shibata T and Himeno A 2000 Flat-passband interleave filter with 200 GHz channel spacing based on planar lightwave circuit-type lattice structure *Electron. Lett.* **36** 1299

[114] Sugita A, Kaneko A, Okamoto K, Itoh M, Himeno A and Ohmori Y 1999 Fabrication of very low insertion loss (~0.8 dB) arrayed-waveguide grating with vertically tapered waveguides *ECOC'99: Proc. 25th European Conf. on Optical Comm.* p 4

[115] Takiguchi K 2001 Recent advances in PLC functional devices *2001 Digest of LEOS Summer Topical Meetings: Advanced Semiconductor Lasers and Applications/Ultraviolet and Blue Lasers and Their Applications/Ultralong Haul DWDM Transmission and Networking/WDM Components*

[116] Khrais N N and Wagner R E 1998 General (de)multiplexer cascade model for transparent digital transmission *J. Opt. Commun.* **19** 75

[117] Amersfoort M and Soole J 1995 Passband flattening of integrated optical filters US 5629992 A1

[118] Vieira Segatto M E, Maxwell G D, Kashyap R and Taylor J R 2001 High-speed transmission and dispersion characteristics of an arrayed-waveguide grating *Opt. Commun. 40* **195** 151

[119] Dragone C 1994 Frequency routing device having a wide and substantially flat passband US 5412744 A1

[120] Rigny A, Bruno A and Sik H 1997 Multigrating method for flattened spectral response wavelength multi/demultiplexer *Electron. Lett.* **33** 1701

[121] Trouchet D 1996 Multiplexer/demultiplexer with flattened spectral response EP0816877 A1

[122] Doerr C R, Stulz L W, Cappuzzo M, Laskowski E, Paunescu A, Gomez L, Gates J V, Shunk S, Chandrasekhar S and Kim H 1999 40-channel programmable integrated add–drop with flat through-spectrum *Proc. ECOC'99 25th European Conf. on Optical Communication* p 46

[123] Heise G, Schneider H W and Clemens P C 1998 Optical phased array filter module with passively compensated temperature dependence *24th European Conf. on Optical Commun. ECOC'98 (IEEE Cat No 98TH8398) Proc. of ECOC '98—24th European Conf. on Optical Communication* p 319

[124] Kaneko A, Kamei S, Inoue Y, Takahashi H and Sugita A 2000 Athermal silica-based arrayed-waveguide grating (AWG) multi/demultiplexers with new low loss groove design *Electron. Lett.* **36** 318

[125] Inoue Y, Ohmori Y, Kawachi M, Ando S, Sawada T and Takahashi H 1994 Polarization mode converter with polyimide half waveplate in silica-based planar lightwave circuits *IEEE Photon. Technol. Lett.* **6** 626

[126] Weiner A M and Kan'an A M 1998 Femtosecond pulse shaping for synthesis processing and time-to-space conversion of ultrafast optical waveforms *IEEE J. Sel. Top. Quantum Electron.* **4** 317

[127] Tsuda H, Takenouchi H, Hirano A, Kurokawa T and Okamoto K 2000 Performance analysis of a dispersion compensator using arrayed-waveguide gratings *J. Lightwave Technol.* **18** 1139

[128] Tsuda H, Takenouchi H, Ishii T, Okamoto K, Goh T, Sato K, Hirano A, Kurokawa T and Amano C 1999 Photonic spectral encoder/decoder using an arrayed-waveguide grating for coherent optical code division multiplexing *OSA Trends in Optics and Photonics WDM Components* vol 29 p 206

[129] Godil A A 2002 Diffractive MEMS for optical networks *Electron. Eng. Design* **74** 43

[130] Gorecki C 2001 Recent advances in silicon guided-wave MOEMS: from technology to application *Opto-Electron. Rev.* **9** 248

[131] Tizhi Huang, Huang J, Yueai Liu, Ming Xu, Yatao Yang, Minchun Li, Chongchang Mao and Jung-Chih Chiao 2001 Performance of a liquid-crystal optical harmonic equalizer *OFC 2001: Optical Fiber Communication Conference and Exhibit Technical Digest Postconference Edition Postdeadline Papers* p PD29

[132] Ranalli A R, Scott B A and Kondis J P 1999 Liquid crystal-based wavelength selectable cross-connect *Proc. ECOC'99 25th European Conf. on Opt. Comm.* p 68

[133] Dimmick T E, Kakarantzas G, Birks T A, Diez A and Russell P S J 2000 Compact all-fibre acoustooptic tunable filters with small bandwidth-length product *IEEE Photon. Technol. Lett.* **12** 1210

[134] Yeralan S, Gunther J, Ritums D L, Cid R, Storey J, Ashmead A C and Popovich M M 2001 Switchable Bragg grating devices for telecommunications applications *Proc. SPIE — The Int. Soc. for Opt. Eng.* vol 4291 p 79

[135] Vallon S, Cayrefourcq I, Chevallier P, Landru N, Alibert G, Laborde P, Little J, Ranalli A and Boos N 2001 Tapped delay line dynamic gain flattening filter *2001 Digest of LEOS Summer Topical Meetings: Advanced Semiconductor Lasers and Applications/Ultraviolet and Blue Lasers and Their Applications/Ultralong Haul DWDM Transmission and Networking/WDM Components)*

[136] Sasayama K, Okuno M and Habara K 1991 Coherent optical transversal filter using silica-based waveguides for high-speed signal processing *J. Lightwave Technol.* **9** 1225

[137] Sasayama K, Okuno M and Habara K 1994 Photonic FDM multichannel selector using coherent optical transversal filter *J. Lightwave Technol.* **12** 664

[138] Offrein B J, Horst F, Bona G L, Germann R, Salemink H W M and Beyeler R 2000 Adaptive gain equalizer in high-index-contrast SiON technology *IEEE Photon. Technol. Lett.* **12** 504

[139] Li Y P and Henry C H 1996 Silica-based optical integrated circuits *IEE Proc. Optoelectron.* **143** 263

[140] Doerr C R, Joyner C H and Stulz L W 1998 Integrated WDM dynamic power equalizer with potentially low insertion loss *IEEE Photon. Technol. Lett.* **10** 1443

[141] Doerr C R, Stulz L W, Pafchek R, Gomez L, Cappuzzo M, Paunescu A, Laskowski E, Buhl L, Kim H K and Chandrasekhar S 2000 An automatic 40-wavelength channelized equalizer *IEEE Photon. Technol. Lett.* **12** 1195

[142] Ramaswami R and Sivarajan K N 1997 *Optical Networks* (San Mateo: Morgan Kaufman)

[143] Ouellette F, Cliche J-F and Gagnon S 1994 All-fibre devices for chromatic dispersion compensation based on chirped distributed resonant coupling *J. Lightwave Technol.* **12** 1728

[144] Madsen C K and Lenz G 2000 A multi-channel dispersion slope compensating optical allpass filter *Optical Fiber Communication Conference Technical Digest Postconference Edition Trends in Optics and Photonics (IEEE Cat No 00CH37079)* p 94

[145] Shirasaki M 1997 Chromatic-dispersion compensator using virtually imaged phased array *IEEE Photon. Technol. Lett.* **9** 1598

[146] Takiguchi K, Okamoto K and Goh T 2001 Integrated optic dispersion slope equaliser for N*20 Gbit/s WDM transmission *Electron. Lett.* **37** 701

[147] Madsen C K, Lenz G, Nielsen T N, Bruce A J, Cappuzzo M A and Gomez L T 1999 Integrated optical allpass filters for dispersion compensation *OSA Trends in Optics and Photonics WDM Components*, vol 29 p 142

[148] Madsen C K 2001 Tunable dispersion compensators based on optical allpass filters *Digest of LEOS Summer Topical Meetings: Advanced Semiconductor Lasers and Applications/Ultraviolet and Blue Lasers and Their Applications/Ultralong Haul DWDM Transmission and Networking/WDM Components (IEEE Cat No 01TH8572)*

[149] Noe R, Heidrich H and Hoffmann D 1988 Endless polarization control systems for coherent optics *J. Lightwave Technol.* **6** 1199

[150] Saida T, Takiguchi K, Kuwahara S, Kisaka Y, Miyamoto Y, Hashizume Y, Shibata T and Okamoto K 2002 Planar lightwave circuit polarization-mode dispersion compensator *IEEE Photon. Technol. Lett.* **14** 507

[151] Okuno M, Sugita A, Jinguji K and Kawachi M 1994 Birefringence control of silica waveguides on Si and its application to a polarization-beam splitter/switch *J. Lightwave Technol.* **12** 625

[152] Cotter D, Manning R J, Blow K J, Ellis A D, Kelly A E, Nesset D, Phillips I D, Poustie A J and Rogers D C 1999 Nonlinear optics for high-speed digital information processing *Science* **286** 1523

[153] Kaminow I P and Koch T L (ed) 1997 *Optical Fiber Telecommunications IIIA* (San Diego: Academic)

[154] Bishop D 2000 Silicon micromachines for lightwave networks *Photonics in Switching Topical Meeting OSA Trends in Optics and Photonics Series* vol 32 p 11

[155] Lin Y, Goldstein E L, Lunardi L M and Tkach R W 1999 Optical crossconnects for high-capacity lightwave networks *J. High Speed Networks* **8** 17

[156] Lin L Y and Goldstein E L 2000 MEMS for optical switching *Photonics in Switching Topical Meeting OSA Trends in Optics and Photonics Series* vol 32 p 23

[157] Neilson D T *et al* 2000 Fully provisioned 112*112 micro-mechanical optical crossconnect with 358 Tb/s demonstrated capacity *Optical Fiber Commun. Conf. Technical Digest Postconf. Ed. Trends in Opt. and Photonics (IEEE Cat No 00CH37079)*, vol 37 p 202

[158] Krähenbühl R and Burns W K 2000 Enhanced crosstalk suppression for Ti:LiNbO$_3$ digital optical switches *Photonics in Switching Topical Meeting OSA Trends in Optics and Photonics Series* vol 32 p 160

[159] Antao Chen, Irvin R W, Murphy E J, Grencavich R, Murphy T O and Richards G W 2000 High performance LiNbO$_3$ switches for multiwavelength optical networks *Photonics in Switching Topical Meeting OSA Trends in Optics and Photonics Series* vol 32 p 163

[160] Kirihara T and Inoue H 1996 InP-based optical switch arrays using semiconductor optical amplifiers *Int. J. High Speed Electron. Systems* **7** 85

[161] van Berlo W, Janson M, Lundgren L, Morner A-C, Terlecki J, Gustavsson M, Granestrand P and Svensson P 1995 Polarization-insensitive monolithic 4*4 InGaAsP–InP laser amplifier gate switch matrix *IEEE Photon. Technol. Lett.* **7** 1291

[162] Fouquet J E 2000 Progress in optical cross-connects for circuit-switched applications *Photonics in Switching Topical Meeting OSA Trends in Optics and Photonics Series* vol 32 14

[163] Sakata T, Togo H and Shimokawa F 2000 Reflection-type 2*2 optical waveguide switch using the Goos–Hanchen shift effect *Appl. Phys. Lett.* **76** 2841

[164] Venkatesh S, Haven R, Chen D, Reynolds H L, Harkins G, Close S, Troll M, Fouquet J E, Schroeder D and McGuire P 2001 Recent advances in bubble-actuated cross-connect switches *Technical Digest CLEO/Pacific Rim 2001 4th Pacific Rim Conf. on Lasers and Electro-Optics (Cat No 01TH8557)* p 1

[165] Betty I, Rousina-Webb R and Chi Wu 2000 A robust, low-crosstalk, InGaAs–InP total-internal-reflection switch for optical cross-connect *Photonics in Switching Topical Meeting OSA Trends in Optics and Photonics Series* vol 32 p 5

[166] Noguchi K 1997 Optical multichannel switch composed of liquid-crystal light-modulator arrays and birefringent crystals *Electron. Lett.* **33** 1627

[167] Dorgeuille F, Ambrosy A, Grieshaber W, Pommereau F, Boubal F, Rabaron S, Gaborit F, Guillemot I, Poucheron C, Le Bris J, Blume O, Lauckner J, Luz G, Matthles K, Ruess K, Schilling M, Schneider S, Noire L, Tregoat D and Artigue C 1999 Loss-free 1*4 opto-hybrid space switch based on an array of 4 gain-clamped SOA gates *Proceedings of ECOC'99. 25th European Conf. on Optical Communication* p 176

[168] Sasaki J, Hatakeyama H, Tamanuki T, Kitamura S, Yamaguchi M, Kitamura N, Shimoda T, Kitamura M, Kato T and Itoh M 1998 Hybrid integrated 4*4 optical matrix switch using self-aligned semiconductor optical amplifier gate arrays and silica planar lightwave circuit *Electron. Lett.* **34** 986

[169] Dorgeuille F, Noirie L, Faure J P, Ambrosy A, Rabaron S, Boubal F, Schilling M and Artigue C 2000 1.28 Tbit/s throughput 8*8 optical switch based on arrays of gain-clamped semiconductor optical amplifier gates *Optical Fiber Communication Conf. Technical Digest Postconference Edition. Trends in Optics and Photonics* vol 37 *(IEEE Cat. No. 00CH37079)* p 221

[170] Kasahara R, Yanagisawa M, Sugita A, Ogawa I, Hashimoto T, Suzaki Y and Magari K 1999 Fabrication of compact optical wavelength selector by integrating arrayed-waveguide-gratings and optical gate array on a single PLC platform *Proc. of ECOC'99: 25th European Conf. on Opt. Comm.* p 122

[171] Renaud M, Keller D, Sahri N, Silvestre S, Prieto D, Dorgeuille F, Pommereau F, Emery J Y, Grard E and Mayer H P 2001 SOA-based optical network components *Proc. 51st Electron. Components and Technol. Conf. (Cat. No. 01CH37220) 2001*, p 433

[172] Goh T, Yasu M, Hattori K, Himeno A, Okuno M and Ohmori Y 1998 Low-loss and high-extinction-ratio silica-based strictly nonblocking 16*16 thermooptic matrix switch *IEEE Photon. Technol. Lett.* **10** 810

[173] Kasahara R, Yanagisawa M, Sugita A, Goh T, Yasu M, Himeno A and Matsui S 1999 Low-power consumption silica-based 2*2 thermooptic switch using trenched silicon substrate *IEEE Photon. Technol. Lett.* **11** 1132

[174] Sohma S, Goh T, Okazaki H, Okuno M and Sugita A 2002 Low switching power silica-based super high delta thermo-optic switch with heat insulating grooves *Electron. Lett.* **38** 127

[175] Hida Y, Onose H and Imamura S 1993 Polymer waveguide thermooptic switch with low electric power consumption at 1.3 μm *IEEE Photon. Technol. Lett.* **5** 782

[176] Murphy E J 1997 Photonics switching *Optical Fiber Telecommunications III* ed B I P Kaminow and T L Koch (New York: Academic) p 463

[177] Doerr C R, Joyner C H, Stulz L W and Monnard R 1998 Wavelength-division multiplexing cross connect in InP *IEEE Photon. Technol. Lett.* **10** 117

[178] Vinchant J-F, Cavailles J A, Erman M, Jarry P and Renaud M 1992 InP/GaInAsP guided-wave phase modulators based on carrier-induced effects: theory and experiment *J. Lightwave Technol.* **10** 63

[179] Nakamura S, Tajima K and Sugimoto Y 1995 High-repetition operation of a symmetric Mach-Zehnder all-optical switch *Appl. Phys. Lett.* **66** 2457

[180] Vreeburg C G M, Uitterdijk T, Oei Y S, Smit M K, Groen F H, Metaal E G, Demeester P and Frankena H J 1997 First InP-based reconfigurable integrated add–drop multiplexer *IEEE Photon. Technol. Lett.* **9** 188

[181] Little B E, Foresi J S, Steinmeyer G, Thoen E R, Chu S T, Hans H A, Ippen E P, Kimerling L C and Greene W 1998 Ultra-compact Si–SiO$_2$ microring resonator optical channel dropping filters *IEEE Photon. Technol. Lett.* **10** 549

[182] Wehrmann F, Harizi C, Herrmann H, Rust U, Sohler W and Westenhofer S 1996 Integrated optical wavelength selective acoustically tunable 2*2 switches (add–drop multiplexers) in $LiNbO_3$ *IEEE J. Sel. Top. Quantum Electron.* **2** 263

[183] Okamoto K, Okuno M, Himeno A and Ohmori Y 1996 16-channel optical add/drop multiplexer consisting of arrayed-waveguide gratings and double-gate switches *Electron. Lett.* **32** 1471

[184] Doerr C R 1998 Proposed WDM cross connect using a planar arrangement of waveguide grating routers and phase shifters *IEEE Photon. Technol. Lett.* **10** 528

[185] Teshima M, Koga M and Sato K 1995 Multiwavelength simultaneous monitoring circuit employing wavelength crossover properties of arrayed-waveguide grating *Electron. Lett.* **31** 1595

[186] Janz C, Dagens B, Emery J-Y, Renaud M and Lavigne B 2000 Integrated SOA-based interferometers for all-optical signal processing *Proc. 26th European Conf. on Opt. Commun.* p 115

[187] Fischer S, Duelk M, Puleo M, Girardi R, Gamper E, Vogt W, Hunziker W, Gini E and Melchior H 1999 40-Gb/s OTDM to 4*10 Gb/s WDM conversion in monolithic InP Mach–Zehnder interferometer module *IEEE Photon. Technol. Lett.* **11** 1262

[188] Ueno Y, Nakamura S, Hatakeyama H, Tamanuki T, Sasaki T and Tajima K 2000 168-Gb/s OTDM wavelength conversion using an SMZ-type all-optical switch *Proc. of 26th European Conf. Opt. Comm.* p 13

[189] Kazovski L, Benedetto S and Willner A E 1996 *Optical Fiber Communication Systems* (Boston: Artech)

[190] Wolfson D, Hansen P B, Kloch A, Fjelde T, Janz C, Coquelin A, Guillemot I, Garorit F, Poingt F and Renaud M 1999 All-optical 2R regeneration at 40 Gbit/s in an SOA-based Mach–Zehnder interferometer *OFC/IOOC'99: Optical Fiber Communication Conf. and Int. Conf. on Integrated Opt. and Optical Fiber Communications (Cat. No 99CH36322)* p PD 36

[191] Mirasso C R *et al* 1999 Self-pulsating semiconductor lasers: theory and experiment *IEEE J. Quantum Electron.* **35** 764

[192] Rees P, McEvoy P, Valle A, O'Gorman J, Lynch S, Landais P, Pesquera L and Hegarty J 1999 A theoretical analysis of optical clock extraction using a self-pulsating laser diode *IEEE J. Quantum Electron.* **35** 221

[193] Barnsley P E, Wickes H J, Wickens G E and Spirit D M 1991 All-optical clock recovery from 5 Gb/s RZ data using a self-pulsating 1.56 μm laser diode *IEEE Photon. Technol. Lett.* **3** 942

[194] Pender W A, Widdowson T and Ellis A D 1996 Error free operation of a 40 Gbit/s all-optical regenerator *Electron. Lett.* **32** 567

[195] Phillips I D, Ellis D, Thiele J, Manning R J and Kelly A E 1998 40 Gbit/s all-optical data regeneration and demultiplexing with long pattern lengths using a semiconductor nonlinear interferometer *Electron. Lett.* **34** 2340

[196] Vallon S 2002 Private communiation

[197] Bourdon G 2002 Private communication

[198] Alibert G 2002 Private communication

[199] Delprat D 2002 Private communication

[200] Lin L Y, Goldstein E L and Tkach R W 1998 Free-space micromachined optical switches with submillisecond switching time for large-scale optical crossconnects *IEEE Photon. Technol. Lett.* **10** 525

Further reading

The textbooks in [80], [22], [20], [13] and [2] develop the theory of waveguides, couplers and splitters as well as electro-optic and magneto-optic control. For approximation methods in channel waveguides and numerical techniques see in particular references [1] and [2].

Material properties of nonlinear materials are found in [8], with details on nonlinear semiconductor devices in [7 chapter 11] and applications of nonlinear devices for TDM systems in [63 chapter 9].

Okamoto [2] and Madsen [22] treat authoritatively single mode planar waveguide building blocks and integrated devices, and [22] is the key reference for the design of optical filters. [20] and [2] have chapters on BPM, and [7 chapter 12] focuses on numeric design tools and their accuracy.

The recent books in [63] and [7] have specific chapters on hybrid and monolithic integration technique as well as lithium niobate components, rare-earth doped glass waveguides and integrated InP devices.

Packaging of opto-electronic devices is discussed in [79] in specific chapters on laser packaging, optical interconnection techniques and interconnection loss budgets.

Optical network components, propagation in fibre and network architectures are described from a practical perspective in [142], and detailed theoretical treatment of (single and multi-channel) signal propagation in fibre and noise in systems can be found in [189]. Finally, Agrawal [106] treats nonlinear effects in fibre and applications such as Raman amplification and nonlinear optic fibre devices.

B8
Infrared devices and techniques

Antoni Rogalski and Krzysztof Chrzanowski

B8.1 Introduction

Looking back over the past 1000 years, we notice that infrared (IR) radiation itself was unknown until 200 years ago when Herschel's experiment with a thermometer was first reported. He built a crude monochromator that used a thermometer as a detector so that he could measure the distribution of energy in sunlight [1]. Following the works of Kirchhoff, Stefan, Boltzmann, Wien and Rayleigh, Max Planck culminated the effort with the well-known Planck's law.

Traditionally, IR technologies are connected with controlling functions and night-vision problems with earlier applications connected simply with detection of IR radiation, and later by forming IR images from temperature and emissivity differences (systems for recognition and surveillance, tank sight systems, anti-tank missiles, air–air missiles). The years during World War II saw the origins of modern IR techniques. Recent success in applying IR technology to remote sensing problems has been made possible by the successful development of high-performance IR detectors over five decades. Most of the funding has been provided to fulfil military needs, but peaceful applications have increased continuously, especially in the last decade of the 20th century. These include medical, industry, earth resources and energy conservation applications. Medical applications include thermography in which IR scans of the body detect cancers or other traumas, which raise the body surface temperature. Earth resource determinations are done by using IR images from satellites in conjunction with field observation for calibration (in this manner, e.g. the area and content of fields and forests can be determined). In some cases, even the state of health of a crop be determined from space. Energy conservation in homes and industry has been aided by the use of IR scans to determine the points of maximum heat loss. Demands to use these technologies are quickly growing due to their effective applications, e.g. in global monitoring of environmental pollution and climate changes, long time prognoses of agriculture crop yield, chemical process monitoring, Fourier transform IR spectrometry, IR astronomy, car driving, IR imaging in medical diagnostics and others.

Nowadays, only about 10% of the market is commercial. After a decade, the commercial market can grow to over 70% in volume and 40% in value, largely connected with volume production of uncooled imagers for automobile driving [2]. In large volume production for automobile drivers, the cost of uncooled imaging systems will decrease to below $1000.

The infrared range covers all electromagnetic radiation longer than the visible, but shorter than millimetre waves. Many proposals of division of the IR range have been published. The division shown below (table B8.1) is used by the military community and is based on the limits of spectral bands of commonly used IR detectors. Wavelength of 1 μm is a sensitivity limit of popular Si detectors. Similarly, wavelength of 3 μm is a long wavelength sensitivity of PbS and InGaAs detectors; wavelength of 6 μm is a sensitivity limit of InSb, PbSe, PtSi detectors and HgCdTe detectors optimized for the 3–5 μm

Table B8.1. Division of infrared radiation.

Region (abbreviation)	Wavelength range (μm)
Near infrared (NIR)	0.78–1
Short wavelength IR (SWIR)	1–3
Medium wavelength IR (MWIR)	3–6
Long wavelength IR (LWIR)	6–15
Very long wavelength IR (VLWIR)	15–30
Far infrared (FIR)	30–100
Submillimetre (SubMM)	100–1000

atmospheric window; and finally wavelength of 15 μm is a long wavelength sensitivity limit of HgCdTe detectors optimized for the 8–14 μm atmospheric window.

B8.2 Infrared system fundamentals

B8.2.1 Thermal emission

All objects are composed of continually vibrating atoms, with higher energy atoms vibrating more frequently. The vibration of all charged particles, including these atoms, generates electromagnetic waves. The higher the temperature of an object, the faster the vibration, and thus the higher the spectral radiant energy. As a result, all objects are continually emitting radiation at a rate with a wavelength distribution that depends upon the temperature of the object and its spectral emissivity, $\varepsilon(\lambda)$.

Radiant emission is usually treated in terms of the concept of a blackbody [3]. A blackbody is an object that absorbs all incident radiation and conversely, according to the Kirchhoff law, is a perfect radiator. The energy emitted by a blackbody is the maximum theoretically possible for a given temperature. The radiative power (or number of photons emitted) and its wavelength distribution are given by the Planck radiation law:

$$W(\lambda, T) = \frac{2\pi hc^2}{\lambda^5}\left[\exp\left(\frac{hc}{\lambda kT}\right) - 1\right]^{-1} \quad \mathrm{W\,cm^{-2}\,\mu m^{-1}} \tag{B8.1}$$

$$P(\lambda, T) = \frac{2\pi c}{\lambda^4}\left[\exp\left(\frac{hc}{\lambda kT}\right) - 1\right]^{-1} \quad \mathrm{photons\,s^{-1}\,cm^{-2}\,\mu m^{-1}} \tag{B8.2}$$

where λ is the wavelength, T the temperature, h Planck's constant, c the velocity of light and k Boltzmann's constant.

Figure B8.1 shows a plot of these curves for a number of blackbody temperatures. As the temperature increases, the amount of energy emitted at any wavelength increases too, and the wavelength of peak emission decreases. The latter is given by the Wien displacement law:

$$\lambda_{\mathrm{mw}} T = 2898\ \mu\mathrm{m\,K} \quad \text{for maximum watts}$$
$$\lambda_{\mathrm{mp}} T = 3670\ \mu\mathrm{m\,K} \quad \text{for maximum photons}$$

The loci of these maxima are shown in figure B8.1. Note that for an object at an ambient temperature of 290 K, λ_{mw} and λ_{mp} occur at 10.0 and 12.7 μm, respectively. We need detectors operating

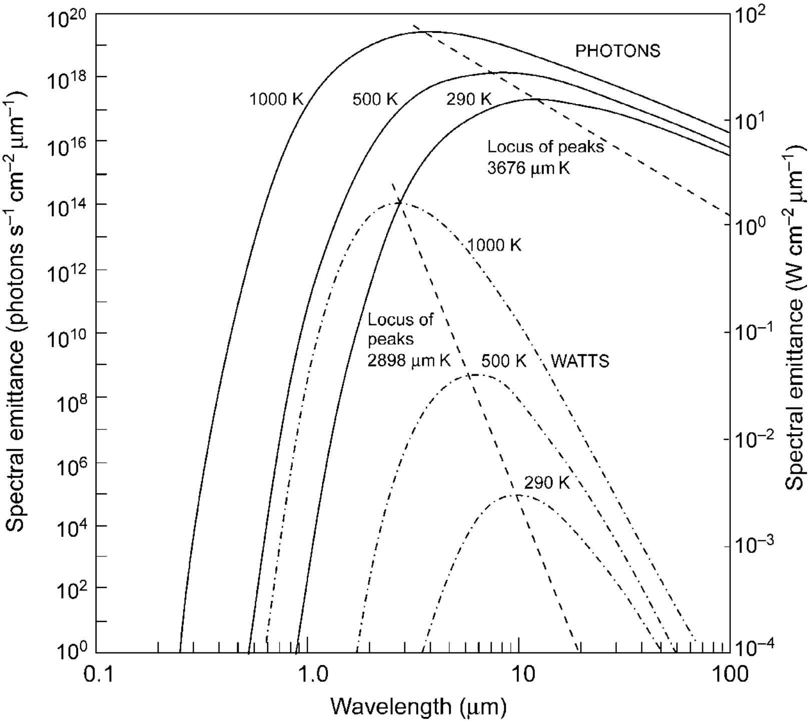


Figure B8.1. Planck's law for spectral emittance. Reproduced from [4].

near 10 μm if we expect to 'see' room temperature objects such as people, trees and trucks without the aid of reflected light. For hotter objects such as engines, maximum emission occurs at shorter wavelengths. Thus, the waveband 2–15 μm in infrared or thermal region of the electromagnetic spectrum contains the maximum radiative emission for thermal imaging purposes.

B8.2.2 Atmospheric transmission

Most of the above mentioned applications require transmission through air, but the radiation is attenuated by the processes of scattering and absorption. Scattering causes a change in the direction of a radiation beam; it is caused by absorption and subsequent reradiation of energy by suspended particles. For larger particles, scattering is independent of wavelength. However, for small particles, compared with the wavelength of the radiation, the process is known as Rayleigh scattering and exhibits a λ^{-4} dependence. Therefore, scattering by gas molecules is negligibly small for wavelengths longer than 2 μm. Also smoke and light mist particles are usually small with respect to IR wavelengths, and IR radiation can therefore penetrate further through smoke and mists than visible radiation. However, rain, fog particles and aerosols are larger and consequently scatter IR and visible radiation to a similar degree.

Figure B8.2 is a plot of the transmission through 6000 ft of air as a function of wavelength. Specific absorption bands of water, carbon dioxide and oxygen molecules are indicated that restrict atmospheric transmission to two windows at 3–5 and 8–14 μm. Ozone, nitrous oxide, carbon monoxide and methane are less important IR absorbing constituents of the atmosphere.

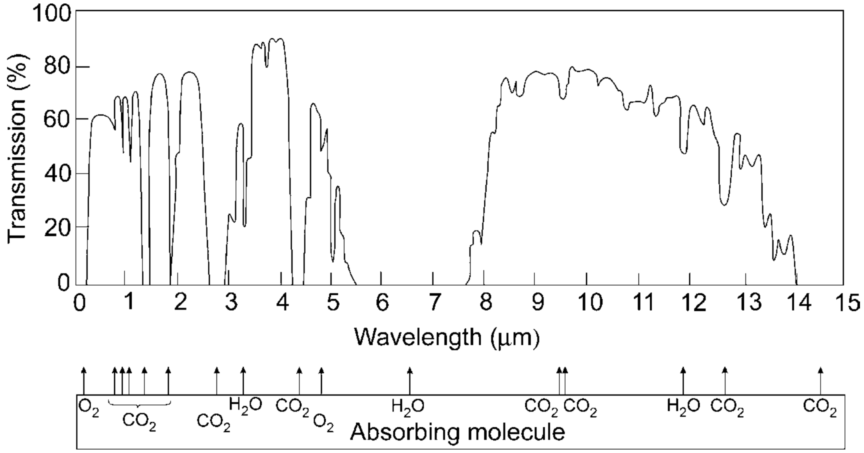


Figure B8.2. Transmission of the atmosphere for a 6000 ft horizontal path at sea level containing 17 mm of precipitate water. Reproduced from [5].

B8.2.3 Scene radiation and contrast

The total radiation received from any object is the sum of the emitted, reflected and transmitted radiation. Objects that are not blackbodies emit only the fraction $\varepsilon(\lambda)$ of blackbody radiation, and the remaining fraction, $1-\varepsilon(\lambda)$, is either transmitted or, for opaque objects, reflected. When the scene is composed of objects and backgrounds of similar temperatures, reflected radiation tends to reduce the available contrast. However, reflections of hotter or colder objects have a significant effect on the appearance of a thermal scene. The powers of 290 K blackbody emission and ground-level solar radiation in MWIR and LWIR bands are given in table B8.2. We can see that while reflected sunlight has a negligible effect on 8–13 μm imaging, it is important in the 3–5 μm band.

A thermal image arises from temperature variations or differences in emissivity within a scene. The thermal contrast is one of the important parameters for IR imaging devices. It is the ratio of the derivative of spectral photon incidence to the spectral photon incidence

$$C = \frac{\partial W/\partial T}{W}$$

The contrast in a thermal image is small when compared with visible image contrast due to differences in reflectivity. For a 291 K object in a 290 K scene, it is about 0.039 in the 3–5 μm band and 0.017 in

Table B8.2. Power available in each MWIR and LWIR imaging band.

IR region (μm)	Ground-level solar radiation (W m^{-2})	Emission from 290 K blackbody (W m^{-2})
3–5	24	4.1
8–13	1.5	127

Reproduced from [4].

the 8–13 μm band. Thus, while the LWIR band may have the higher sensitivity for ambient temperature objects, the MWIR band has the greater contrast.

B8.2.4 Choice of IR band

In general, the 8–14 μm band is preferred for high performance thermal imaging because of its higher sensitivity to ambient temperature objects and its better transmission through mist and smoke. However, the 3–5 μm band may be more appropriate for hotter objects, or if sensitivity is less important than contrast. Also additional differences occur, e.g. the advantage of the MWIR band is the smaller diameter of the optics required to obtain a certain resolution and that some detectors may operate at higher temperatures (thermoelectric cooling) than is usual in the LWIR band where cryogenic cooling is required (about 77 K).

Summarizing, MWIR and LWIR μm spectral bands differ substantially with respect to background flux, scene characteristics, temperature contrast and atmospheric transmission under diverse weather conditions. Factors that favour MWIR applications are: higher contrast, superior clear-weather performance (favourable weather conditions, e.g. in most countries of Asia and Africa), higher transmittivity in high humidity and higher resolution due to ~3 × smaller optical diffraction. Factors that favour LWIR applications are: better performance in fog and dust conditions, winter haze (typical weather conditions, e.g. in West Europe, North USA, Canada), higher immunity to atmospheric turbulence and reduced sensitivity to solar glints and fire flares. The possibility of achieving higher signal-to-noise (S/N) ratio due to greater radiance levels in LWIR spectral range is not persuasive because the background photon fluxes are higher to the same extent, and also because of readout limitation possibilities. Theoretically, in staring arrays charge can be integrated for the full frame time, but because of restrictions in the charge-handling capacity of the readout cells, it is much less compared to the frame time, especially for LWIR detectors for which background photon flux exceeds the useful signals by orders of magnitude.

B8.2.5 Detectors

The figure of merit used for detectors is detectivity. It has been found in many instances that this parameter varies inversely with the square root of both the detector's sensitive area, A, and the electrical bandwidth, Δf. In order to simplify the comparison of different detectors, the following definition has been introduced [6]

$$D^* = \frac{(A\Delta f)^{1/2}}{\Phi_e}(\mathrm{SNR}) \tag{B8.3}$$

where Φ_e is the spectral radiant incident power. D^* is defined as the rms signal-to-noise ratio (SNR) in a 1 Hz bandwidth per unit rms incident radiation power per square root of detector area. D^* is expressed in cm Hz$^{1/2}$ W^{-1}, which is recently called 'Jones'. Spectral detectivity curves for a number of commercially available IR detectors are shown in figure B8.3. Interest has centred mainly on the wavelengths of the two atmospheric windows 3–5 and 8–14 μm, though in recent years there has been increasing interest in longer wavelengths stimulated by space applications.

Progress in IR detector technology is connected mainly to semiconductor IR detectors, which are included in the class of photon detectors. In the class of photon detectors, the radiation is absorbed within the material by interaction with electrons. The observed electrical output signal results from the changed electronic energy distribution. The photon detectors show a selective wavelength dependence of the response per unit incident radiation power. They exhibit both perfect signal-to-noise performance and a very fast response. But to achieve this, the photon detectors require cryogenic cooling.

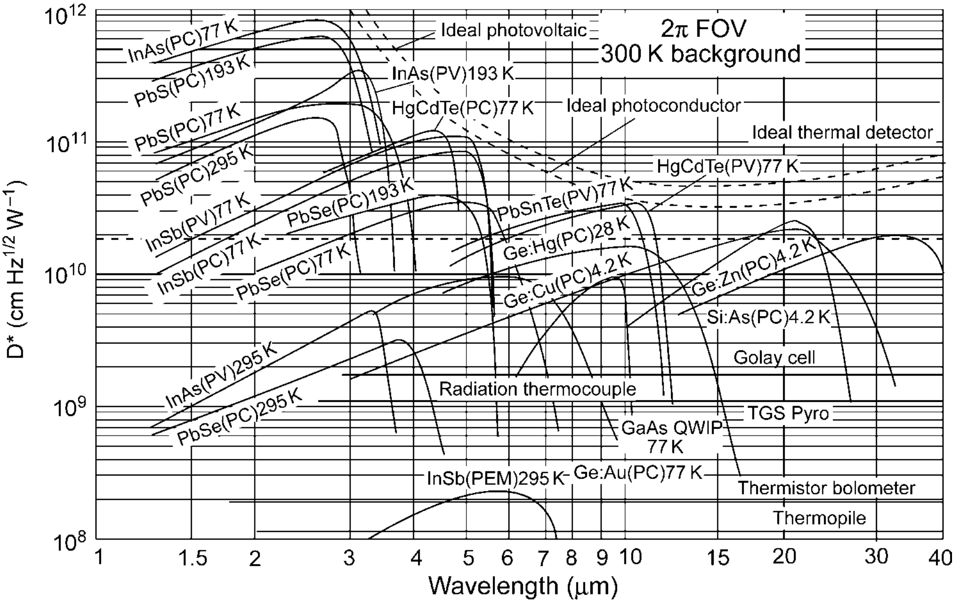


Figure B8.3. Comparison of the D^* of various commercially available infrared detectors when operated at the indicated temperature. The chopping frequency is 1000 Hz for all detectors except the thermopile (10 Hz), thermocouple (10 Hz), thermistor bolometer (10 Hz), Golay cell (10 Hz) and pyroelectric detector (10 Hz). Each detector is assumed to view a hemispherical surround at a temperature of 300 K. Theoretical curves for the background-limited D^* for ideal photovoltaic and photoconductive detectors and thermal detectors are also shown.

Cooling requirements are the main obstacle to the more widespread use of IR systems based on semiconductor photodetectors making them bulky, heavy, expensive and inconvenient to use. Depending on the nature of interaction, the class of photon detectors is further sub-divided into different types. The most important are: intrinsic detectors (HgCdTe, InGaAs, InSb, PbS, PbSe), extrinsic detectors (Si:As, Si:Ga), photoemissive (metal silicide Schottky barriers) detectors and quantum well detectors (GaAs/AlGaAs QWIPs).

The second class of IR detectors is composed of thermal detectors. In a thermal detector, the incident radiation is absorbed to change the temperature of material, and the resultant change in some physical properties is used to generate an electrical output. The detector element is suspended on lags, which are connected to the heat sink. Thermal effects are generally wavelength independent; the signal depends upon the radiant power (or its rate of change) but not upon its spectral content. In pyroelectric detectors, a change in the internal spontaneous polarization is measured, whereas in the case of bolometers a change in the electrical resistance is measured. In contrast to photon detectors, thermal detectors typically operate at room temperature. They are usually characterized by modest sensitivity and slow response but they are cheap and easy to use. The greatest utility in IR technology has been found by bolometers, pyroelectric detectors and thermopiles. Typical values of detectivities of thermal detectors at 10 Hz change in the range between 10^8 and 10^9 cm Hz$^{1/2}$ W^{-1}.

Up until the nineties of the 20th century, thermal detectors have been considerably less exploited in commercial and military systems in comparison with photon detectors. The reason for this disparity is that thermal detectors are popularly believed to be rather slow and insensitive in comparison with photon detectors. As a result, the worldwide effort to develop thermal detectors was extremely small relative to that of photon detectors. In the last decade, however, it has been shown that extremely good imagery can be obtained from large thermal detector arrays operating uncooled at TV frame rates. The speed of thermal detectors is quite adequate for nonscanned imagers with two-dimensional (2D) detectors. The moderate sensitivity of thermal detectors can be compensated by a large number of elements in 2D electronically scanned arrays. With large arrays of thermal detectors the best values of NEDT, below 0.1 K, could be reached because effective noise bandwidths less than 100 Hz can be achieved.

B8.2.6 Cooling

The signal output of a photon detector is so small that at ordinary temperatures it is swamped by the thermal noise due to random generation and recombination of carriers in the semiconductor. In order to reduce the thermal generation of carriers and minimize noise, photon detectors must be cooled and must therefore be encapsulated. The method of cooling varies according to the operating temperature and the system's logistical requirements. Most 8–14 μm detectors operate at about 77 K and can be cooled by liquid nitrogen. In the field, however, it is more convenient to use compressed air and a Joule–Thompson minicooler [7]. The operation of the Joule–Thompson cooler is based on the fact that as the high-pressure gas expands on leaving a throttle valve, it cools and liquefies. The gas used must be purified to remove water vapour and carbon dioxide which could freeze and block the throttle valve. Specially designed Joule–Thompson coolers using argon are suitable for ultra-fast cool-down.

The use of cooling engines, in particular those employing the Stirling cycle [8], has increased recently due to their efficiency, reliability and cost reduction. The Stirling engine requires several minutes cool-down time; the working fluid is helium. Both Joule–Thompson and engine-cooled detectors are housed in precision-bore Dewars into which the cooling device is inserted (see figure B8.4). Mounted in the vacuum space at the end of the inner wall of the Dewar, and surrounded by a cooled radiation shield compatible with the convergence angle of the optical system, the detector looks out through an IR window. In some Dewars, the electrical leads to detector elements are embedded in the inner wall of the dewar to protect them from damage due to vibration.

Many detectors in the 3–5 μm waveband are thermoelectrically cooled. In this case, detectors are usually mounted in a hermetic encapsulation with a base designed to make good contact with a heatsink.

B8.2.7 IR optics

The optical block in an IR system creates an image of observed objects in the plane of the detector (detectors). In the case of a scanning imager, the optical scanning system creates an image with the number of pixels much greater than the number of elements of the detector. In addition, optical elements like windows, domes and filters can be used to protect the system from the environment or to modify the detector spectral response.

There is no essential difference in design rules of optical objectives for visible and IR ranges. The designer of IR optics is only more limited because there are significantly fewer materials suitable for IR optical elements, in comparison with those for the visible range, particularly for wavelengths over 2.5 μm.

There are two types of IR optical element: reflective elements and refractive elements. As the names suggest, the role of reflective elements is to reflect incident radiation and the role of refractive elements is to refract and transmit incident radiation.

Mirrors used extensively inside IR systems (especially in scanners) are most often met as reflective elements that serve manifold functions in IR systems. Elsewhere they need a protective coating to prevent them from tarnishing. Spherical or aspherical mirrors are employed as imaging elements. Flat mirrors are widely used to fold optical paths, and reflective prisms are often used in scanning systems.

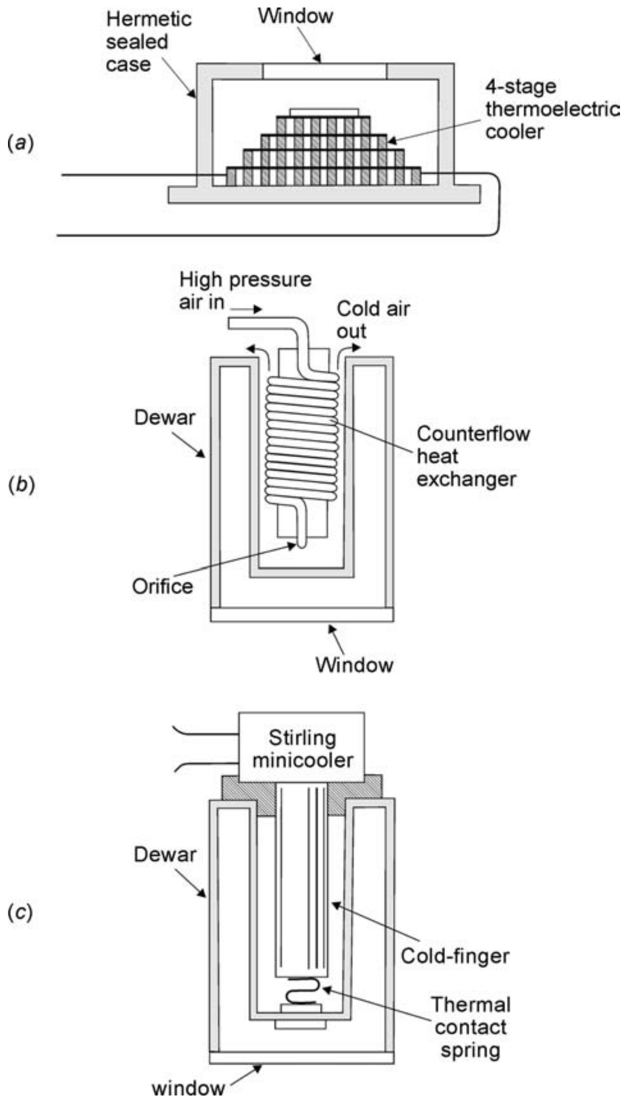

(a)
Hermetic sealed case
Window
4-stage thermoelectric cooler
(b)
High pressure air in
Cold air out
Dewar
Counterflow heat exchanger
Orifice
Window
(c)
Stirling minicooler
Dewar
Cold-finger
Thermal contact spring
window

Figure B8.4. Three ways of cooling IR detectors: (*a*) four-stage thermoelectric cooler (Peltier effect), (*b*) Joule–Thompson cooler and (*c*) Stirling-cycle engine.

Four materials are most often used for mirror fabrication: optical crown glass, low-expansion borosilicate glass (LEBG), synthetic fused silica and Zerodur. Less popular in use are metallic substrates (beryllium, copper) and silicon carbide. Optical crown glass is typically applied in nonimaging systems. It has a relatively high thermal expansion coefficient and is employed when thermal stability is not a critical factor. LEBG, known by the Corning brand name Pyrex, is well suited for high quality front-surface mirrors designed for low optical deformation under thermal shock. Synthetic fused silica has a very low thermal expansion coefficient.

Metallic coatings are typically used as reflective coatings of IR mirrors. There are four types of most often used metallic coating: bare aluminium, protected aluminium, silver and gold. They offer high reflectivity, over about 95%, in the 3–15 μm spectral range. Bare aluminium has a very high reflectance value but oxidizes over time. Protected aluminium is a bare aluminium coating with a dielectric overcoat that arrests the oxidation process. Silver offers better reflectance in the near IR than aluminium and high reflectance across a broad spectrum. Gold is a widely used material and offers consistently very high reflectance (about 99%) in the 0.8–50 μm range. However, gold is soft (it cannot be touched to remove dust) and is most often used in the laboratory.

The most popular materials used in manufacturing refractive optics of IR systems are: germanium (Ge), silicon (Si), fused silica (SiO_2), glass BK-7, zinc selenide (ZnSe) and zinc sulfide (ZnS). The IR-transmitting materials potentially available for use as windows and lenses are gathered in table B8.3 and their IR transmission is shown in figure B8.5.

Germanium is a silvery metallic-appearing solid of very high refractive index (≈ 4) that enables design of high-resolution optical systems using a minimal number of germanium lenses. Its useful

Table B8.3. Principal characteristics of some infrared materials.

Material	Waveband (μm)	$n_{4\,\mu m}$, $n_{10\,\mu m}$	dn/dT ($10^{-6}\,K^{-1}$)	Density ($g\,cm^{-3}$)	Other characteristics
Ge	3–5, 8–12	4.025, 4.004	424 (4 μm), 404 (10 μm)	5.33	Brittle, semiconductor, can be diamond turned, visibly opaque, hard
Si	3–5	3.425	159 (5 μm)	2.33	Brittle, semiconductor, diamond turned with difficulty, visibly opaque, hard
GaAs	3–5, 8–12	3.304, 3.274	150	5.32	Brittle, semiconductor, visibly opaque, hard
ZnS	3–5, 8–12	2.252, 2.200	43 (4 μm), 41 (10 μm)	4.09	Yellowish, moderate hardness and strength, can be diamond turned, scatters short wavelengths
ZnSe	3–5, 8–12	2.433, 2.406	63 (4 μm), 60 (10 μm)	5.26	Yellow–orange, relatively soft and weak, can be diamond turned, very low internal absorption and scatter
CaF_2	3–5	1.410	−8.1 (3.39 μm)	3.18	Visibly clear, can be diamond turned, mildly hygroscopic
Sapphire	3–5	1.677 (n_o) 1.667 (n_e)	6 (o), 12 (e)	3.99	Very hard, difficult to polish due to crystal boundaries
AMTIR-1	3–5, 8–12	2.513, 2.497	72 (10 μm)	4.41	Amorphous IR glass, can be slumped to near-net shape
BK7 (glass)	0.35–2.3		3.4	2.51	Typical optical glass

Reproduced from [9].

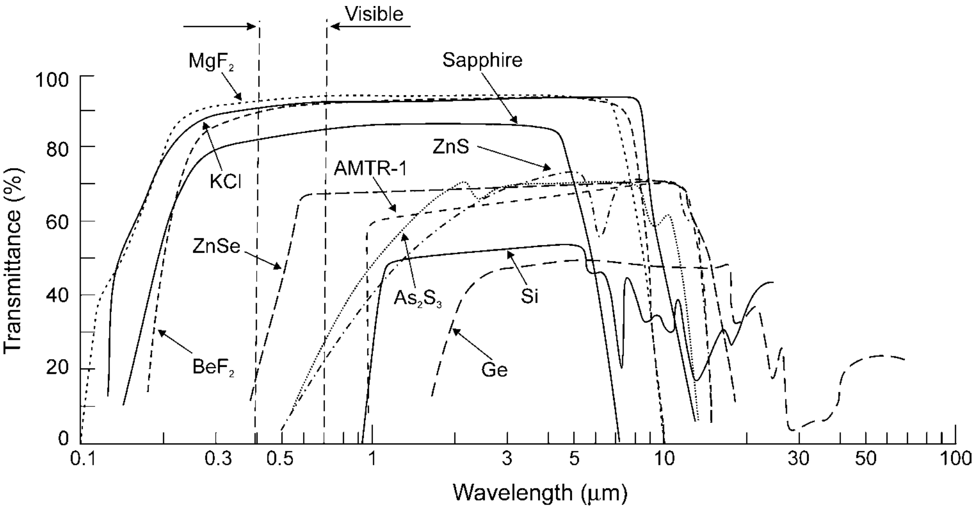


Figure B8.5. Transmission range of infrared materials. Reproduced from [9].

transmission range is from 2 to about 15 μm. It is quite brittle and difficult to cut but accepts a very good polish. Additionally, due to its very high refractive index, antireflection coatings are essential for any germanium transmitting optical system. Germanium has a low dispersion and is unlikely to need colour correcting except in the highest-resolution systems. In spite of high material price and cost of antireflection coatings, germanium lenses are particularly useful for the 8–12 μm band. A significant disadvantage of germanium is the serious dependence of its refractive index on temperature, so germanium telescopes and lenses may need to be athermalized.

Physical and chemical properties of silicon are very similar to properties of germanium. It has high refractive index (≈ 3.45), is brittle, does not cleave, takes an excellent polish and has large dn/dT. Similarly to germanium, silicon optics must have antireflection coatings. Silicon offers two transmission ranges: 1–7 and 25–300 μm. Only the first one is used in typical IR systems. The material is significantly cheaper than germanium. It is used mostly for IR systems operating in the 3–5 μm band.

Single-crystal material has generally higher transmission than polycrystalline material. Optical-grade germanium used for the highest optical transmission is n-type doped to receive a conductivity of 5–14 Ω cm. Silicon is used in its intrinsic state. At elevated temperatures, semiconducting materials become opaque. As a result, germanium is of little use above 100°C. In the 8–14 μm region, semi-insulating GaAs may be used at the temperatures up to 200°C.

Ordinary glass does not transmit radiation beyond 2.5 μm in the IR region. Fused silica is characterized by a very low thermal expansion coefficient that makes optical systems particularly useful in changing environmental conditions. It offers a transmission range from about 0.3 to 3 μm. Because of low reflection losses due to the low refractive index (≈ 1.45), antireflection coatings are not needed. However, an antireflection coating is recommended to avoid ghost images. Fused silica is more expensive than BK-7, but still significantly cheaper than Ge, ZnS and ZnSe, and is a popular material for lenses of IR systems with bands located below 3 μm. BK-7 glass characteristics are similar to fused silica; the difference is only a bit shorter transmission band up to 2.5 μm.

ZnSe is expensive material comparable to germanium; has a transmission range from 2 to about 20 μm, and a refractive index about 2.4. It is partially translucent in the visible and reddish in colour. Due to the relatively high refractive index, antireflection coatings are necessary. The chemical resistance of the material is excellent.

ZnS has excellent transmission in the range from 2 to 12 μm. It is usually a polycrystalline material that shows only a light yellow colour. Recently also colourless, low-scatter grades of ZnS have become available. Because of the relatively high refractive index of 2.25, antireflection coatings are needed to minimize flux reflection. The hardness and fracture strength are very good. ZnS is brittle, can operate at elevated temperatures and also can be used to colour correct high-performance germanium optics.

The alkali halides have excellent IR transmission; however, they are either soft or brittle and many of them are attacked by moisture, making them generally unsuitable for industrial applications. For more detailed discussion of the IR materials, see references [10, 11].

B8.2.8 Night-vision system concepts

Night-vision systems can be divided into two categories: those depending upon the reception and processing of radiation reflected by an object and those which operate with radiation internally generated by an object. The latter systems are described in the subsection below.

The human visual perception system is optimized to operate in daytime illumination conditions. The visual spectrum extends from about 420 to 700 nm and the region of greatest sensitivity is near the peak wavelength of sunlight at around 550 nm. However, at night fewer visible light photons are available and only large, high-contrast objects are visible. It appears that the photon rate in the region from 800 to 900 nm is five to seven times greater than in the visible region around 500 nm. Moreover, the reflectivity of various materials (e.g. green vegetation, because of its chlorophyll content) is higher between 800 and 900 nm than at 500 nm. It means that at night more light is available in the NIR than in the visual region and that against certain backgrounds more contrast is available.

A considerable improvement in night-vision capability can be achieved with night viewing equipment which consists of an objective lens, image intensifier and eyepiece (see figure B8.6). Improved visibility is obtained by gathering more light from the scene with an objective lens than the unaided eye; by use of a photocathode that has higher photosensitivity and broader spectral response than the eye; and by amplification of photo-events for visual sensation.

B8.2.9 Thermal imaging system concepts

Thermal imaging is a technique for converting a scene's thermal radiation pattern (invisible to the human eye) into a visible image. Its usefulness is due to the following aspects [4]:

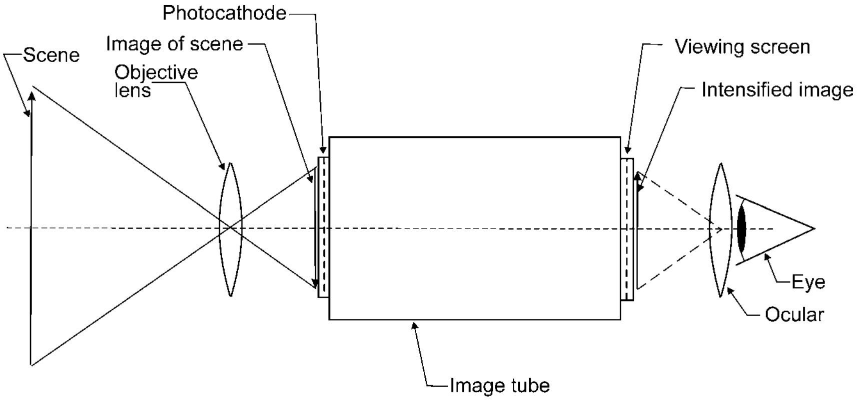


Figure B8.6. Diagram of an image intensifier.

- it is totally passive technique and allows day and night operation;
- it is ideal for detection of hot or cold spots, or areas of different emissivities, within a scene;
- thermal radiation can penetrate smoke and mist more readily than visible radiation;
- it is a real-time, remote sensing technique.

The thermal image is a pictorial representation of temperature difference. Displayed on a scanned raster, the image resembles a television picture of the scene and can be computer processed to colour-code temperature ranges. Originally developed (in the 1960s) to extend the scope of night-vision systems, thermal imagers at first provided an alternative to image intensifiers. As the technology has matured, its range of application has expanded and now extends into the fields that have little or nothing to do with night vision (e.g. stress analysis, medical diagnostics). In most present-day thermal imagers, an optically focused image is scanned (mechanically or electronically) across detectors (many elements or 2D array) the output of which is converted into a visual image. The optics, mode of scanning and signal processing electronics are closely interrelated. The number of picture points in the scene is governed by the nature of the detector (its performance) or the size of the detector array. The effective number of picture points or resolution elements in the scene may be increased by an optomechanical scanning device which images different parts of the scene on to the detector sequentially in time.

The performance of a thermal imager is usually specified in terms of temperature resolution. It can be shown that the temperature sensitivity of an imager, the so-called noise equivalent temperature (NETD), can be given by [12]:

$$\mathrm{NETD} = \frac{4f_{\#}^{2}(\Delta f)^{1/2}}{A^{1/2}t_{\mathrm{op}}M^{*}} \tag{B8.4}$$

where $f_{\#}$ is the f-number of the detector optics ($f_{\#} = f/D$, f is the focal length and D the diameter of the lens), t_{op} the transmission of the optics and M^{*} the figure of merit that includes not only the detector performance D^{*} but also the spectral dependence of the emitted radiation, $(\partial S/\partial T)_{\lambda}$, and the atmospheric transmission t_{at}. It is given by the following equation:

$$M^{*} = \int_{0}^{\infty}\left(\frac{\partial S}{\partial T}\right)_{\lambda} t_{\mathrm{at}\lambda}D_{\lambda}^{*}\,\mathrm{d}\lambda \tag{B8.5}$$

NETD is the difference of temperature of the object required to produce an electric signal equal to the rms noise at the input of the display [6]. Temperature resolution depends on the efficiency of the optical system, responsivity and noise of the detector, and SNR of the signal processing circuitry.

For high sensitivity, the NETD must be low. The sensitivity increases inversely as the square root of the electrical bandwidth. For a given size of IR scene, the electronic bandwidth is inversely proportional to the number of parallel detector elements, and so the thermal sensitivity increases as the square root of the total number of detector elements, irrespective of the parallel or serial content of the array.

B8.3 IR systems

This section briefly concentrates on selected IR systems and is arranged in order of increasing complexity: smart weapon seekers, FLIRs and space-based systems. A comprehensive compendium devoted to IR systems was copublished in 1993 by the Infrared Information Analysis Center (IRIA) and the International Society for Optical Engineering (SPIE) as *The Infrared and Electro-Optical Systems Handbook* (executive editors: Joseph S Accetta and David L Shumaker).

B8.3.1 Smart weapon seekers

The seeker is the primary homing instrument for smart weapons that include missiles, bombs, artillery projectiles and standoff cruise missiles. They can be categorized into three groups: passive nonimaging seekers, passive imaging seekers and active laser guided seekers.

Passive nonimaging seekers use a circular optical plate with adjacent transparent and nontransparent parts called the reticle that is fixed at the image plane of the imaging optics of the missile head (figure B8.7). A single IR detector, of size a bit larger than the reticle, is placed just behind it. The location of the point image of the target on the reticle plate changes, even when the target does not change its position, due to rotation of the reticle or rotation of the imaging optics. Therefore, radiation emitted by the target generates electrical pulses at the detector output. The pulse duration and phase of these pulses give information about the angular position of the target (figure B8.8).

The grandfather of passive IR seekers is the Sidewinder seeker developed in the 1950s: it employed vacuum tubes and a lead salt single-element detector. During the next decades it was found that, despite their simplicity, passive nonimaging seekers are very effective for guiding missiles when the target is on a uniform background. Therefore, at present, the majority of currently used short-range smart missiles use this type of seeker. However, the effectiveness of passive nonimaging seekers decreases significantly for

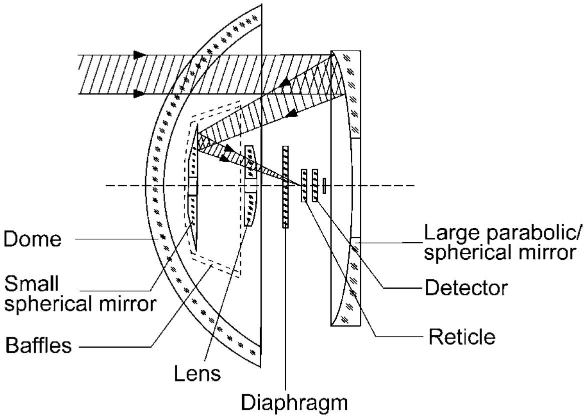


Figure B8.7. Optical diagram of a typical passive nonimaging seeker.

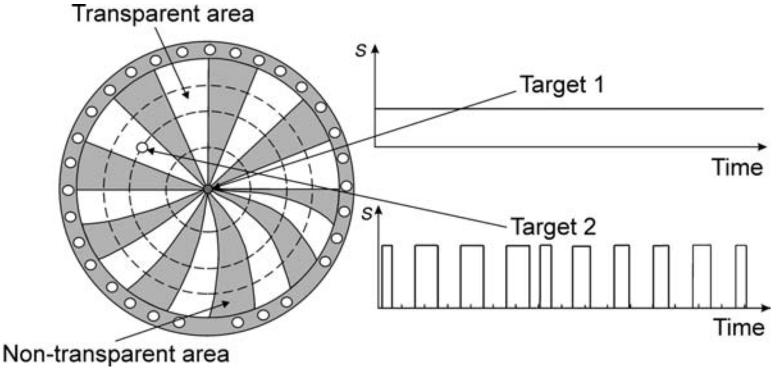


Figure B8.8. Exemplary reticle and the signal generated at the detector output by a few targets of different location.

targets on nonuniform background like typical ground military targets or in the presence of countermeasures. Therefore, the trend of future systems is towards passive imaging seekers.

Passive imaging seekers have a TV camera or a thermal camera in their optoelectronic head. The location of a target is determined from analysis of the image generated by a camera. Some of the air-to-ground missiles attack and destroy ground targets, particularly large nonmovable targets like bridges, bunkers, buildings, etc. However, significant technical limitations exist. First, the seekers using TV cameras can operate only in daylight conditions. Second, it is very difficult to design a thermal camera for high-speed missiles. Such a camera must be of small size, very fast operating, reliable, ready to withstand harsh environmental requirements and of low manufacturing cost. Therefore, the imaging missiles using thermal cameras in their optical heads are still at a development stage.

Active laser guided seekers can be divided into two subclasses: seekers homing on the irradiated target and seekers irradiated with a laser beam (see figure B8.9). Seekers homing on the irradiated target cooperate with a laser illuminator and use the laser radiation reflected by the target. These seekers enable very accurate location of small targets in a highly nonuniform background and are particularly well suited for air-to-ground missiles or bombs. However, it is an active method and employing warning systems or other countermeasures can significantly reduce its effectiveness.

Seekers irradiated with a laser beam are kept on their flight to the target within the beam emitted by the laser illuminator that irradiates the target. Laser radiation, which gives information on target location, comes directly from the illuminator to the sensors at the back of the missile, not after the reflection by the target as in the previous method. Therefore, low-power illuminators can be used here and the effectiveness of the warning systems is reduced.

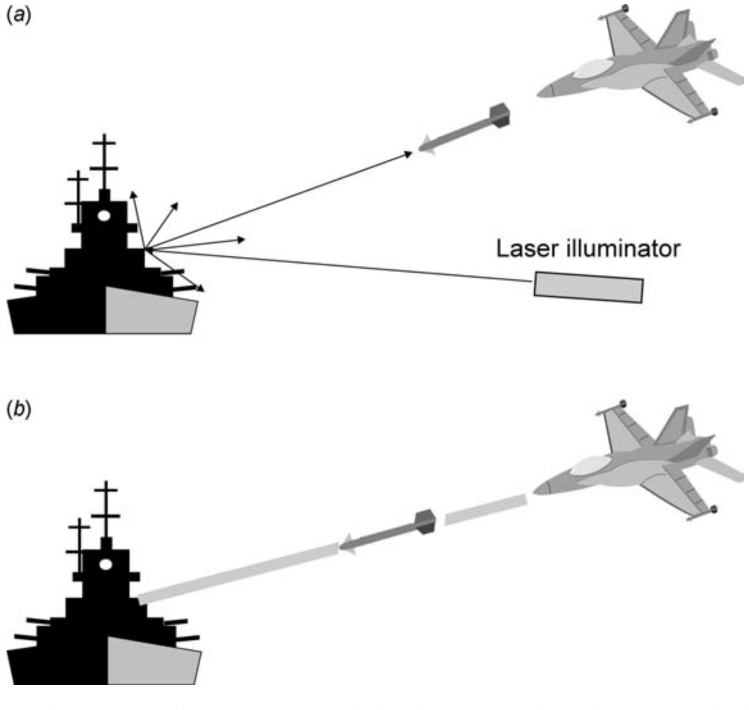


Figure B8.9. Principle of operation of active laser guided seekers: (*a*) seekers homing on the irradiated target and (*b*) seekers irradiated with a laser beam.

The new generation of standoff weapons relies on real-time target recognition, discrimination, tracking, navigation and night vision. It is predicted that the smart weapons will tend to replace the radar emphasis as stealth platforms are increasingly used for low-intensity conflicts. It is more difficult to perform IR missile warning than radar guided missile warning.

A representative architecture of a staring seeker is shown in figure B8.10. To keep seeker volume, weight and power requirements low, only the minimum hardware needed to sense the scene is included. We can note that the seeker's output is going to a missile-based processor, behind the FPA in the seeker, to perform tracking and aimpoint selection. The concept of the seeker's operation includes a standby turn-on, followed by a commit, which cools the FPA. At the beginning, the seeker is locked onto its target by an external sensor or a human. Next, the missile is launched and flies out locked onto its target, matching any target movement. Finally, when the target is close and imaged, the missile chooses an aimpoint and conducts final manoeuvres to get to the target or selects a point and time to fuse and explode.

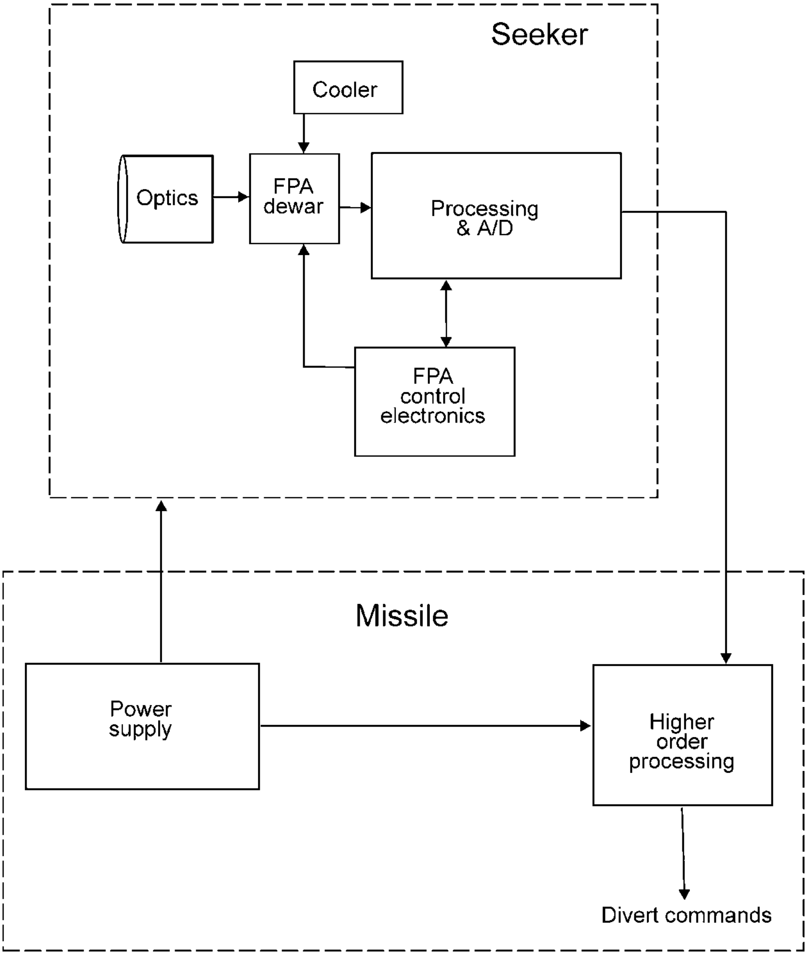


Figure B8.10. Representative imaging (staring) seeker architecture. Reproduced from [7].

B8.3.2 Image intensifier systems

The image intensifiers are classed by generation (Gen) numbers. Gen0 refers to the technology of World War II, employing fragile, vacuum-enveloped photon detectors with poor sensitivity and little gain. Gen1 represents the technology of the early Vietnam era, the 1960s. In this era, the first passive systems, able to amplify ambient starlight, were introduced. Though sensitive, these devices were large and heavy. Gen1 devices used tri-alkali photocathodes to achieve gain of about 1000. By the early 1970s, the microchannel-plate (MCP) amplifier was developed comprising more than two million microscopic conducting channels of hollow glass, each of about 10 μm in diameter, fused into a disc-shaped array. Coupling the MCP with multi-alkali photocathodes, capable of emitting more electrons per incident photon, produced GenII. GenII devices boasted amplifications of 20 000 and operational lives of 2500 h. Interim improvements in bias voltage and construction methods produced the GenII+ version. Substantial improvements in gain and bandwidth in the 1980s heralded the advent of GenIII. Gallium arsenide photocathodes and internal changes in the MCP design resulted in gains ranging from 30 000 to 50 000 and operating lives of 10 000 h.

Many candidate technologies could form the basis of a GenIV, ranging from enhanced current designs to completely different approaches. Among those are devices with a new photocathode that extend spectral response to 1.6 μm and the use of an amplifying mechanism other than MCPs. Other potentials include lightweight systems that fuse the outputs from image intensifiers and thermal imagers, and those that couple electron-bombarded CCD arrays—providing sensitivity in the NIR and MWIR regions—with miniature flat-panel displays. The first GenIV tubes demonstrated substantial increase in target detection range and resolution, particularly at extremely low light levels.

Figure B8.11 shows the response of a typical Gen3 image intensifier superimposed on the night sky radiation spectrum. This figure also shows the CIE photopic curve illustrating the spectral response of the human visual perception system, and the GenII response.

Various implementations of image intensifier tubes have been realized. Phosphor output image intensifiers were reviewed in depth by Csorba [14]. The image is focused onto a semitransparent photocathode and photoelectrons are emitted with a spatial intensity distribution which matches the focused image. In image intensifiers, the electrons are then accelerated towards a phosphor screen where they reproduce the original image with enhanced intensity. Three common forms of image tube are shown in figure B8.12.

In a ‘proximity-focused’ tube, a high electric field (typically 5 kV), and a short distance between the photocathode and the screen, limit spreading of electrons to preserve an image. This form of tube is compact, the image is free from distortion and only a simple power supply is required. However, the resolution of such a tube is limited by the field strength at the photocathode and the resolution is highest when the distance between cathode and screen is small.

An electrostatically focused tube is based upon a system of concentric spheres (cathode and anode, typical bias voltage of 15 kV). In practice, the electrodes depart radically from the simple spherical concept. Additional electrodes can be introduced to provide focusing control and reduce the image distortion, while fibre-optic windows at input and output can be used to improve image quality and provide a better matching to objective and coupling optics. Power suppliers are very simple and lightweight, so this type of tube is widely used in portable applications.

A magnetically focused system gives very high-resolution imagers with little or no distortion. The focusing coil, however, is usually heavy and power consuming. For the best picture quality, the power suppliers for both tube and coil must be stable. This type of tube is used in applications where resolution and low distortion are vital and weight and power consumption do not create unacceptable problems.

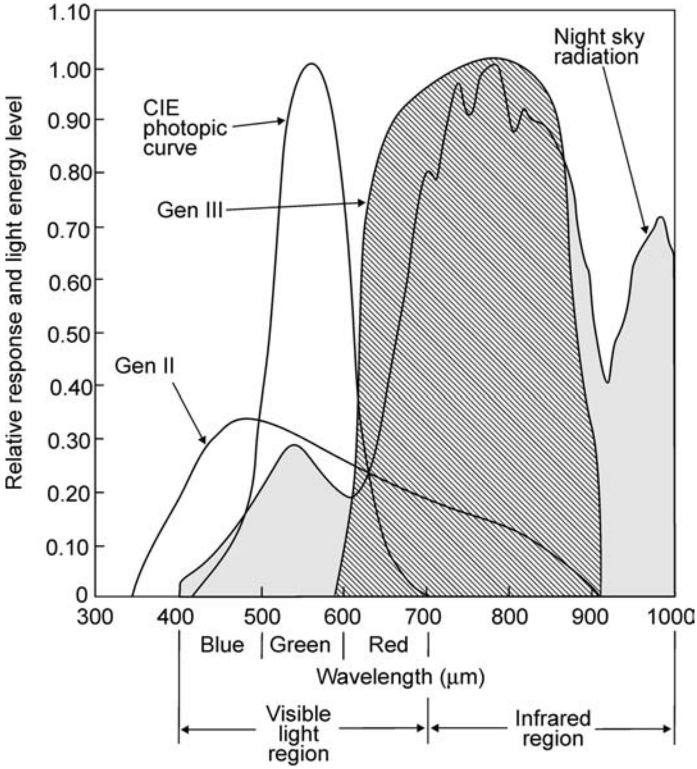


Figure B8.11. Image intensifier tube spectral response curves. Reproduced from [13].

Image intensifiers are widespread in many military applications [15]. The advent of night-vision devices and helmet-mounted displays places additional constraints on the helmet, which is now an important element of the cockpit display system, providing weapon aiming, and other information—such as aircraft attitude and status—to the pilot. For example, figure B8.13 illustrates Marconi Avionics' conventional product produced in large quantities.

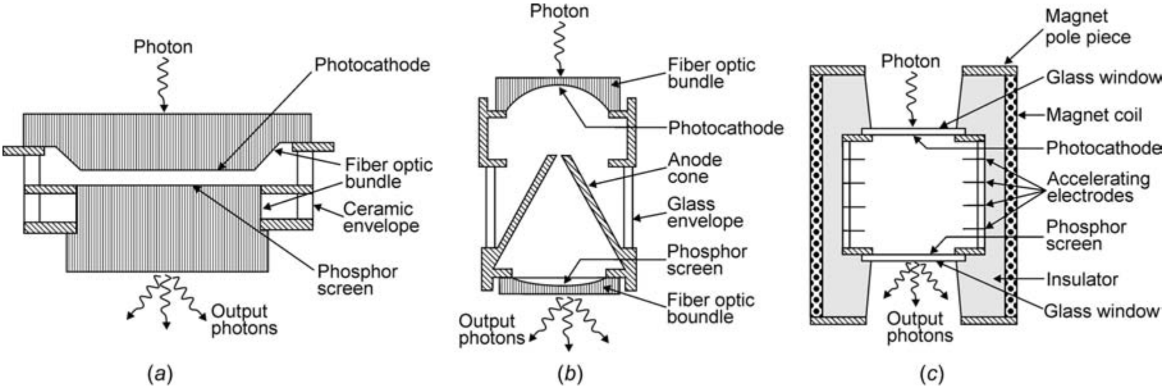

Photon
Photocathode
Fiber optic bundle
Ceramic envelope
Phosphor screen
Output photons
(a)
Photon
Fiber optic bundle
Photocathode
Anode cone
Glass envelope
Phosphor screen
Fiber optic boundle
Output photons
(b)
Magnet pole piece
Photon
Glass window
Magnet coil
Photocathode
Accelerating electrodes
Phosphor screen
Insulator
Glass window
Output photons
(c)

Figure B8.12. Cross-sectional diagrams of a variety of image intensifier types: (*a*) proximity focused, (*b*) electronically focused and (*c*) magnetically focused. Reproduced from [14].

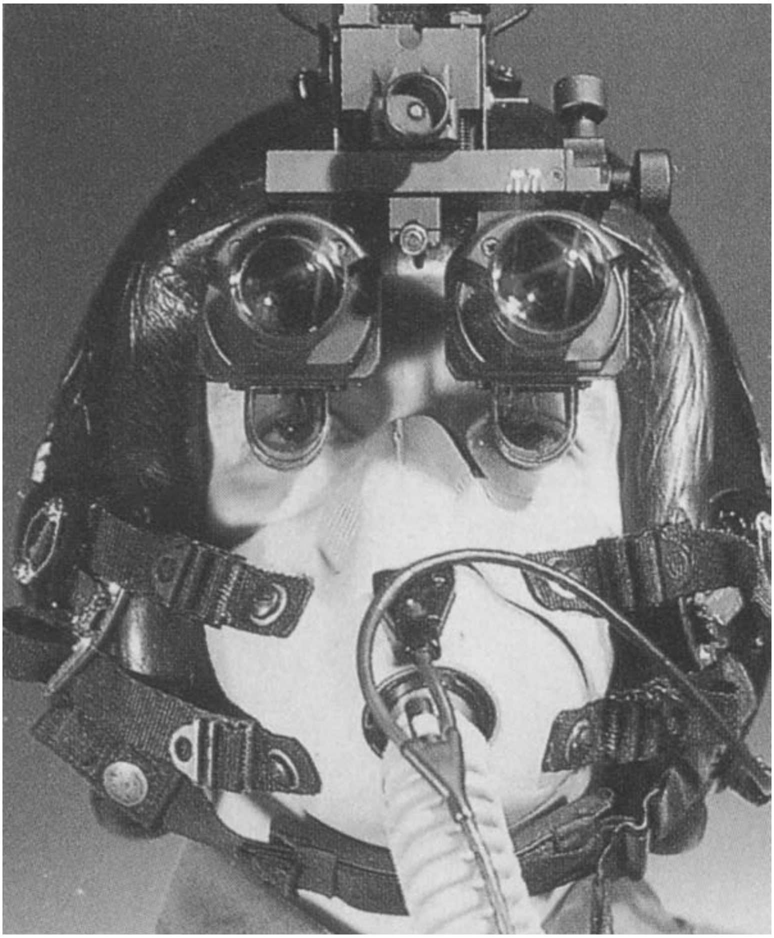


Figure B8.13. Marconi Avionics' Cats Eyes incorporate an optical combiner assembly for each eye, allowing the pilot to view the cockpit and the outside world directly with the night imagery superimposed on it. Cats Eyes have a 30° FOV and weight 820 g, including the helmet plate. Reproduced from [16].

B8.3.3 Thermal imaging systems

The basic concept of a modern thermal imager system is to form a real image of the IR scene, detect the variation in the imaged radiation, and, by suitable electronic processing, create a visible representation of this variation analogous to conventional television cameras.

Due to existing terminology confusion in the literature, we can find at least 11 different terms used as synonyms of the earlier defined thermal imaging systems: thermal imager, thermal camera, thermal imaging camera, FLIR (forward looking infrared), IR imaging system, thermograph, thermovision, thermal viewer, infrared viewer, infrared imaging radiometer, thermal viewer, thermal data viewer and thermal video system. The only real difference between the above-mentioned terms is that the designations 'thermograph', 'infrared imaging radiometer' and 'thermovision' usually refer to thermal cameras used for measurement applications, while the other terms refer to thermal cameras used in observation applications. For example, thermographic imagers supply quantitative temperature, while radiometers provide quantitative radiometric data on the scene (such as radiance or irradiance) or process these data to yield information about temperatures.

Thermal imagers have various applications, depending on the platform and user. Most of them are used in military applications. They often have multiple fields of view (FOV) that are user switchable during operation, which gives both a wide, general surveillance mode as well as a high magnification and narrow field for targeting, designating or detailed intelligence gathering. Many military thermal imagers are integrated with a TV camera and a laser range finder. A TV colour camera is used during daytime conditions due to its superior image quality. Nonmilitary uses include generic search and track, snow rescue, mountain rescue, illegal border crossing detection and pilot assistance at night or in bad weather, forest fire detection, fire fighting, inspection and discreet surveillance and evidence gathering. A small but increasing group of thermal imagers enables noncontact temperature measurement and these cameras are used in areas of industry, science and medicine.

The simplest scanning linear array used in thermal imaging systems, the so-called focal plane array (FPA), consists of a row of detectors (figure B8.14(*a*)). An image is generated by scanning the scene across the strip using, as a rule, a mechanical scanner. At standard video frame rates, at each pixel (detector) a short integration time has been applied and the total charge is accommodated. A staring array is a 2D array of detector pixels (figure B8.14(*b*)) scanned electronically.

The scanning system, which does not include multiplexing functions in the focal plane, belongs to the first-generation systems. A typical example of this kind of detector is a linear photoconductive array (PbS, PbSe, HgCdTe) in which an electrical contact for each element of a multielement array is brought off the cryogenically cooled focal plane to the outside, where one electronic channel is used at ambient

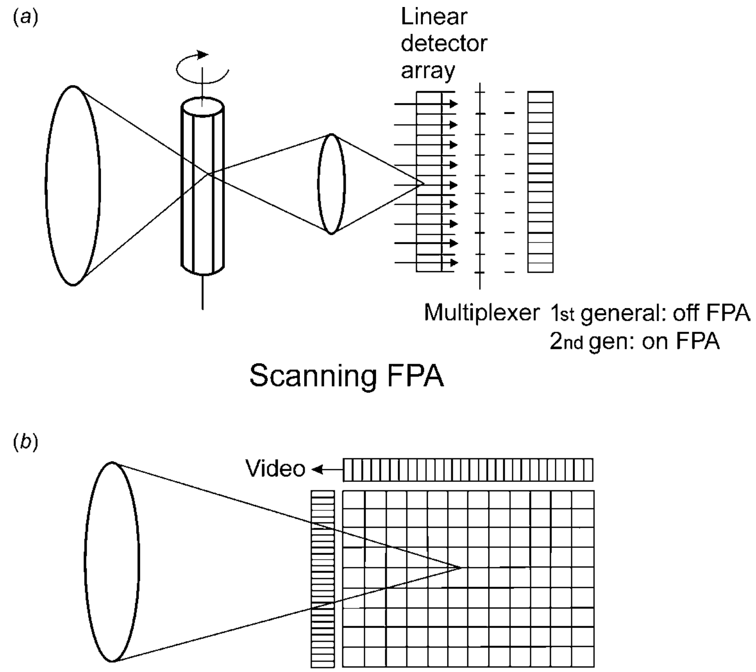


Figure B8.14. Scanning and staring focal plane arrays.

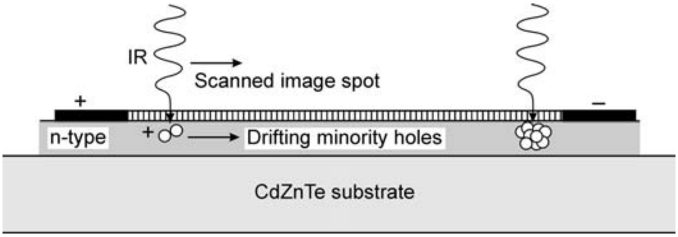


Figure B8.15. Cross-section of a SPRITE photoconductive detector.

temperature for each detector element. The US common module HgCdTe arrays employ 60, 120 or 180 photoconductive elements depending on the application.

A novel variation of the standard photoconductive device, the SPRITE detector, was invented in England [17]. A family of thermal imaging systems has utilized this device; however, now decline in its usage is observed. The SPRITE detector provides signal averaging of a scanned image spot which is accomplished by synchronization between the drift velocity of minority carriers along the length of a photoconductive bar of material and the scan velocity of the imaging system (see figure B8.15). Then the image signal builds up a bundle of minority charge which is collected at the end of the photoconductive bar, effectively integrating the signal for a significant length of time and thereby improving the signal-to-noise ratio.

The second-generation systems (full-framing systems), being developed at present, have at least three orders of magnitude more elements ($>10^6$) on the focal plane than first-generation systems and the detector elements are configured in a 2D array. These staring arrays are scanned electronically by circuits integrated with the arrays. These readout integrated circuits (ROICs) include, e.g. pixel deselecting, antiblooming on each pixel, subframe imaging, output preamplifiers and some other functions. The optics merely focuses the IR image onto the matrix of sensitive elements.

Intermediary systems are also fabricated with multiplexed scanned photodetector linear arrays in use and with, as a rule, time delay and integration (TDI) functions. Typical examples of these systems are HgCdTe multilinear 288×4 arrays fabricated by Sofradir, both for 3–5 and 8–10.5 μm bands with signal processing in the focal plane (photocurrent integration, skimming, partitioning, TDI function, output preamplification and some others).

A number of architectures are used in development of IR FPAs [18]. In general, they may be classified as hybrid and monolithic ones, but these distinctions are often not as important as proponents and critics state them to be. The central design questions involve performance advantages versus ultimate producibility. Each application may favour a different approach depending on the technical requirements, projected costs and schedule.

In the monolithic approach (see figure B8.16), some of the multiplexing is done in the detector material itself rather than in an external readout circuit. The basic element of a monolithic array is a metal–insulator–semiconductor (MIS) structure. An MIS capacitor detects and integrates the IR-generated photocurrent. Although efforts have been made to develop monolithic FPAs using narrow-gap semiconductors, silicon-based FPA technology with Schottky-barrier detectors is the only technology matured to a level of practical use.

Hybrid FPA detectors and multiplexers are fabricated on different substrates and mated with each other by flip-chip bonding (figure B8.17) or loophole interconnection. In this case, we can optimize the detector material and multiplexer independently. Other advantages of the hybrid FPAs are near 100% fill factor and increased signal-processing area on the multiplexer chip. In the flip-chip bonding, the detector array is typically connected by pressure contacts via indium bumps to the silicon multiplex

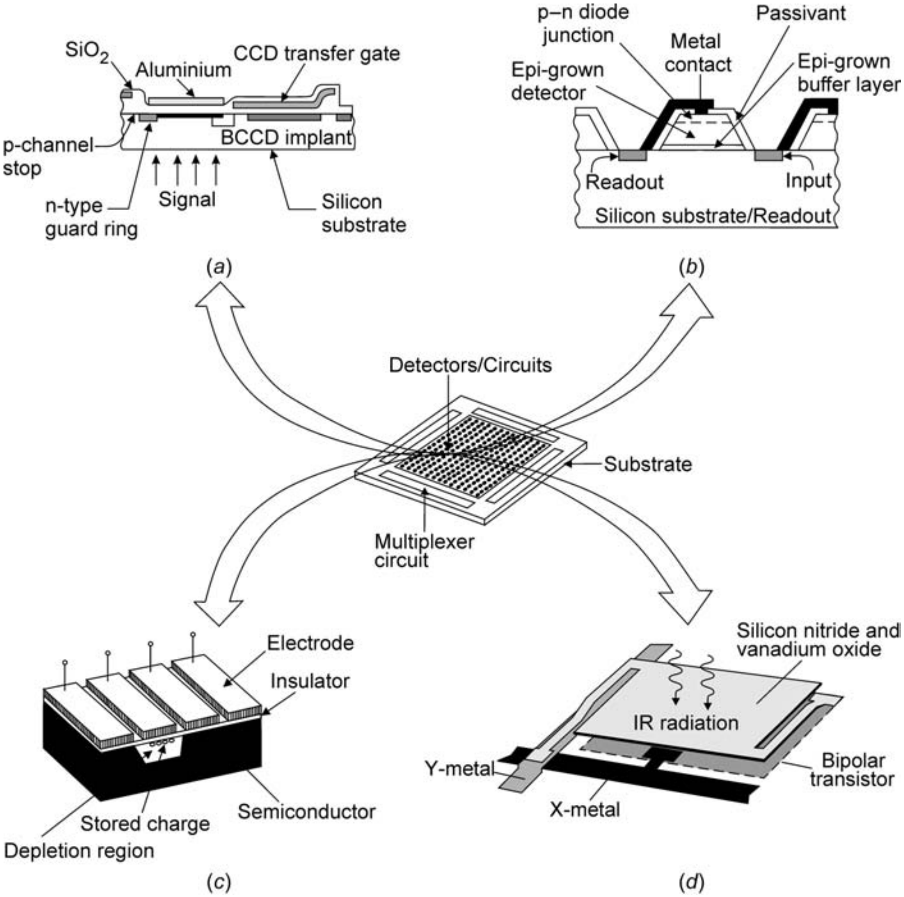

SiO$_2$
Aluminium
CCD transfer gate
p-channel stop
BCCD implant
n-type guard ring
Signal
Silicon substrate
(a)
p–n diode junction
Metal contact
Passivant
Epi-grown detector
Epi-grown buffer layer
Readout
Input
Silicon substrate/Readout
(b)
Detectors/Circuits
Substrate
Multiplexer circuit
Electrode
Insulator
Semiconductor
Stored charge
Depletion region
(c)
Silicon nitride and vanadium oxide
IR radiation
Y-metal
X-metal
Bipolar transistor
(d)

Figure B8.16. Monolithic IR FPAs: (*a*) all silicon, (*b*) heteroepitaxy-on-silicon, (*c*) non-silicon (e.g. HgCdTe CCD) and (*d*) microbolometer.

pads. The detector array can be illuminated from either the frontside or backside (with photons passing through the transparent detector array substrate). In general, the latter approach is most advantageous. When using opaque materials, substrates must be thinned to 10–20 μm in order to obtain sufficient quantum efficiencies and reduce the crosstalk.

Two types of silicon addressing circuit have been developed: CCDs and complementary metal–oxide–semiconductor (CMOS) switches. In CCD addressing circuits, the photogenerated carriers are first integrated in the well formed by a photogate and subsequently transferred to slow (vertical) and fast (horizontal) CCD shift registers [19].

An attractive alternative to the CCD readout is coordinative addressing with CMOS switches. The advantages of CMOS are the existing foundries. Design rules of 0.25 μm are in production with pre-production runs of the 0.13 μm design rules. At present, CMOS with a minimum feature ≤ 0.5 μm is also enabling monolithic visible CMOS imagers.

The minimum resolvable temperature difference (MRTD) is currently considered as the most important parameter of thermal imaging systems (see reference [20]). MRTD enables us to estimate the probability of detection, recognition and identification of military targets knowing the MRTD of

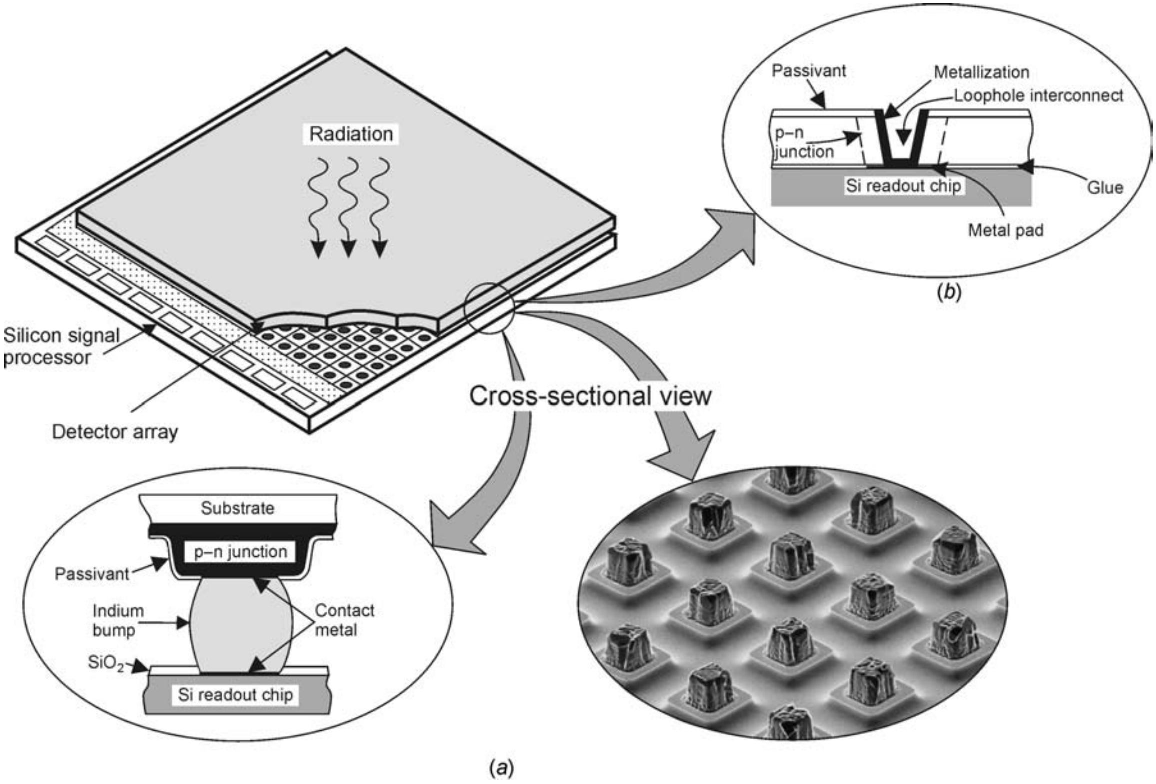

Radiation
Silicon signal processor
Detector array
Cross-sectional view
Passivant
Metallization
Loophole interconnect
p–n junction
Si readout chip
Glue
Metal pad
(b)
Substrate
p–n junction
Passivant
Indium bump
Contact metal
SiO2
Si readout chip
(a)

Figure B8.17. Hybrid IR FPA interconnect techniques between a detector array and silicon multiplexer: (*a*) indium bump technique and (*b*) loophole technique.

the evaluated thermal imager. Military standards determining testing of the thermal imaging systems usually specify that MRTD values for a set of spatial frequencies of the tested imager must be lower than certain values if the imager is to pass the test.

The MRTD is a subjective parameter that describes ability of the imager–human system for detection of low-contrast details of the tested object. It is measured as a minimum temperature difference between the bars of the standard four-bar target and the background required to resolve the thermal image of the bars by an observer versus spatial frequency of the target (see figure B8.18). The measurement results of typical military thermal imagers for airborne surveillance are shown in figure B8.19.

IR cameras versus FLIR systems

Historically, a 'camera' includes neither the storage medium nor the display, while the 'camera system' includes the complete package. At present, the manufacturers offer an optional recording medium (usually CD-ROM), display, and electronics for the display. For example, figure B8.20 is a photograph of the Inframetrics microbolometer IR camera ThermaCam 395.

'FLIR' is archaic 1960s jargon for forward-looking IR to distinguish these systems from IR line scanners, which look down rather than forward. Conversely, most sensors that do look forward are not considered to be FLIRs (e.g. cameras and astronomical instruments). The term

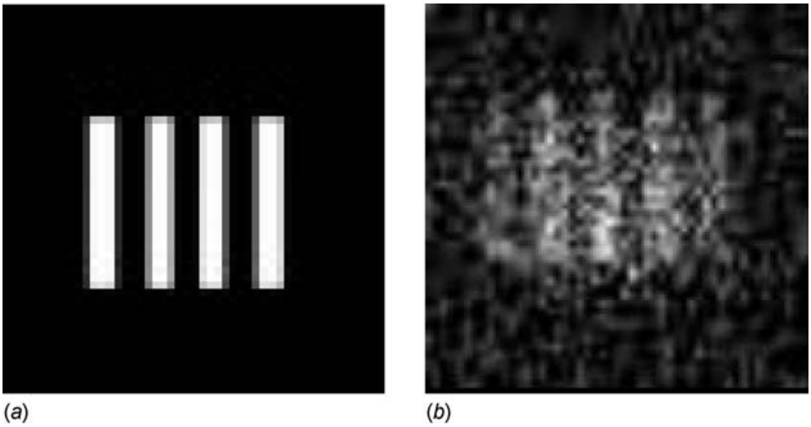


Figure B8.18. Image of standard four-bar target used during MRTD measurement: (*a*) high temperature difference between target and background and (*b*) low temperature difference between target and background.

'FLIR' should be eliminated from IR techno-speak, but is still used and is likely to remain in the jargon for a while.

It is difficult to explain the difference between camera and an FLIR system. In general, FLIRs are designed for specific applications and specific platforms, their optics is integrated into the package, and they are used mostly by people. Cameras usually rely on 'imaging' of a 'target' and they are designed for generic purposes, without much consideration for form and fit; they can be used with many different fore optics and are often used by computers and machines (not just people).

The term 'FLIR' usually implies military or paramilitary use, air-based units and scanners. The FLIR provides automatic search, acquisition, tracking, precision navigation and weapon delivery functions. A typical FLIR is comprised of four line replaceable units, such as an FLIR

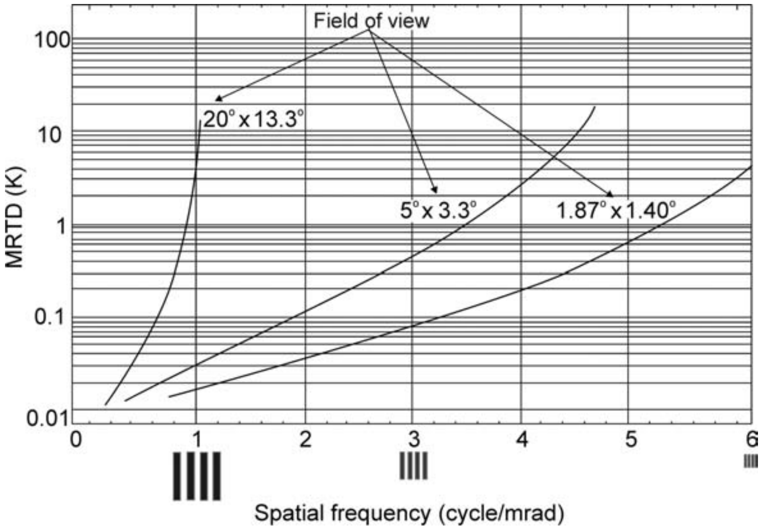


Figure B8.19. MRTD of exemplary military thermal camera used in airborne surveillance.

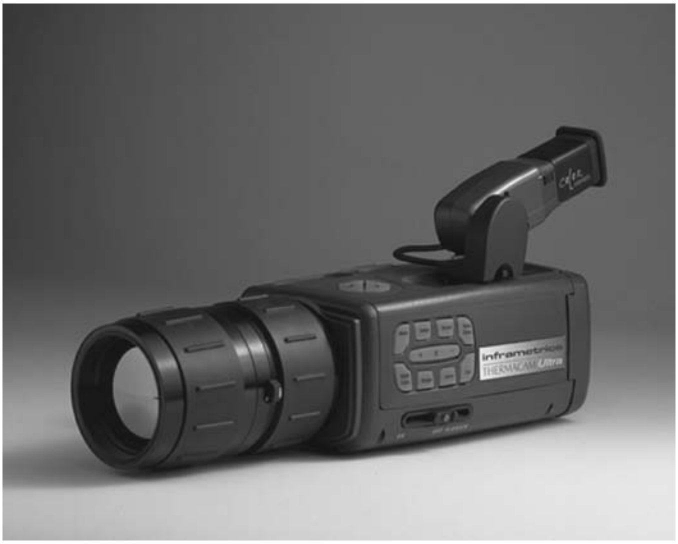


Figure B8.20. Inframetrics microbolometer IR camera ThermaCam 395 (photography courtesy of Inframetrics).

optical assembly mounted on a gyro-stabilized platform, an electronics module containing all necessary electronics circuits and a cryogenically cooled detector array, a power supply unit, and a control and processing assembly.

Figure B8.19 shows representative camera architecture with three distinct hardware pieces: a camera head (which contains optics, including collecting, imaging, zoom, focusing and spectral filtering assemblies), an electronics/control processing box and the display. Electronics and motors to control and drive moving parts must be included. The control electronics usually consist of communication circuits, bias generators and clocks. Usually the camera's sensor (FPA) needs cooling and therefore some form of cooler is included, along with its closed-loop cooling control electronics. The signal from the FPA is of low voltage and amperage and requires analogue preprocessing (including amplification, control and correction), which is located physically near the FPA and included in the camera head. Often, the A/D is also included here. For user convenience, the camera head often contains the minimum hardware needed to keep volume, weight and power to a minimum.

Typical costs of cryogenically cooled imagers around \$50 000 restrict their installation to critical military applications allowing conducting of operations in complete darkness. Moving from cooled to uncooled operation (e.g. using silicon microbolometer or BST pyroelectric arrays) reduces the cost of an imager to below \$20 000. The cost of pyroelectric vidicons is usually a few thousand dollars; however, they have NEDT of 0.5°C (although some of them are now reported as low as 0.2°C) and typically poor image quality compared with full-frame staring arrays. They present a major departure from the camera architecture presented in figure B8.21.

Cameras usually produce high-quality imagers with NEDTs of 0.05–0.1°C. Details and resolution vary by optics and focal planes. A good camera produces an image akin to that of a black and white television.

In the 1960s the earliest FLIRs were linear scanners. In the 1970s first-generation common modules (including a Dewar containing 60, 120 or 180 discrete elements of photoconductive HgCdTe) were introduced. The next generation of FLIRs employed a dense linear array of photovoltaic HgCdTe,

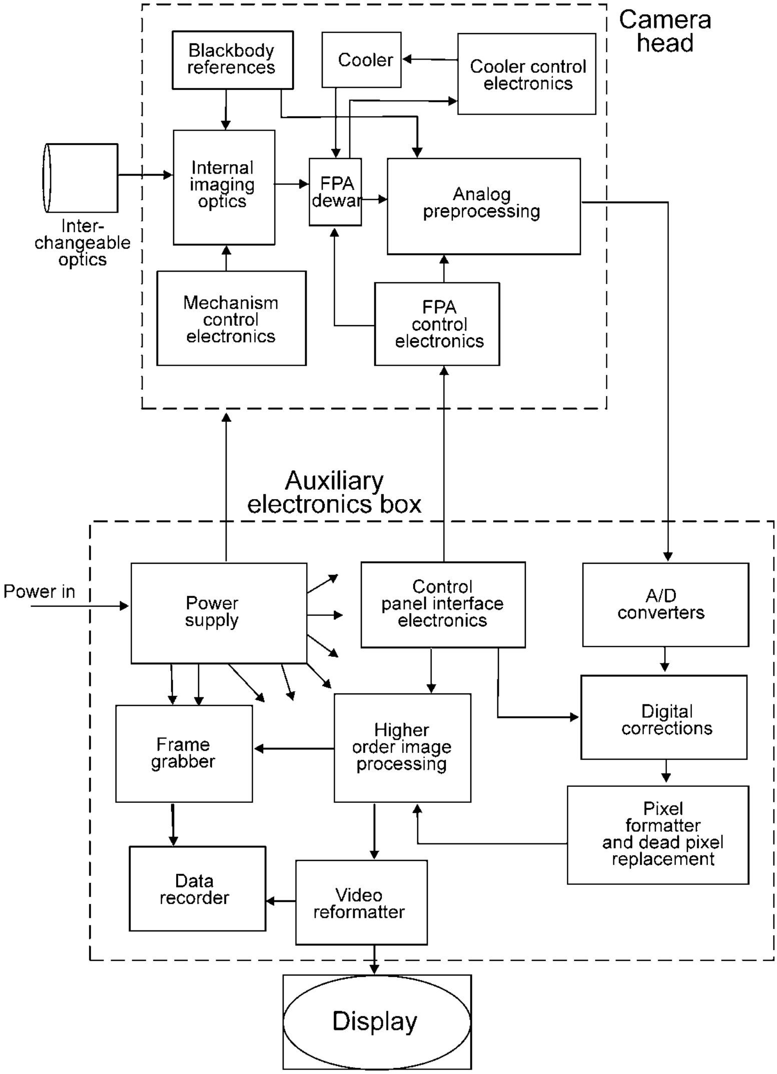

Camera head
Blackbody references
Cooler
Cooler control electronics
Inter-changeable optics
Internal imaging optics
FPA dewar
Analog preprocessing
Mechanism control electronics
FPA control electronics
Auxiliary electronics box
Power in
Power supply
Control panel interface electronics
A/D converters
Digital corrections
Frame grabber
Higher order image processing
Pixel formatter and dead pixel replacement
Data recorder
Video reformatter
Display

Figure B8.21. Representative IR camera system architecture. Reproduced from [7].

usually 2(4) × 480 or 2(4) × 960 elements in TDI for each element. At present, these systems are replaced by full-framing FLIRs that employ a staring arrays (PtSi, HgCdTe, InSb and QWIP).

FLIRs are usually in several discrete packages referred to as line replaceable units (LRUs) such as: scanner head, power supply, image processor, recorder, display and controls. They have the form of boxes spread around the host platform. The controls and display must be mounted in the cockpit with the humans. A representative FLIR architecture with the video signal output (to support LRU for image and higher-order processing) is shown in figure B8.22. Many systems depart significantly from the architecture of figure B8.22.

FLIRs usually use telescopes in the sense that the lens system is focused at a distance very much larger than the focal length. Characteristics such as field of view (FOV), resolution, element size and

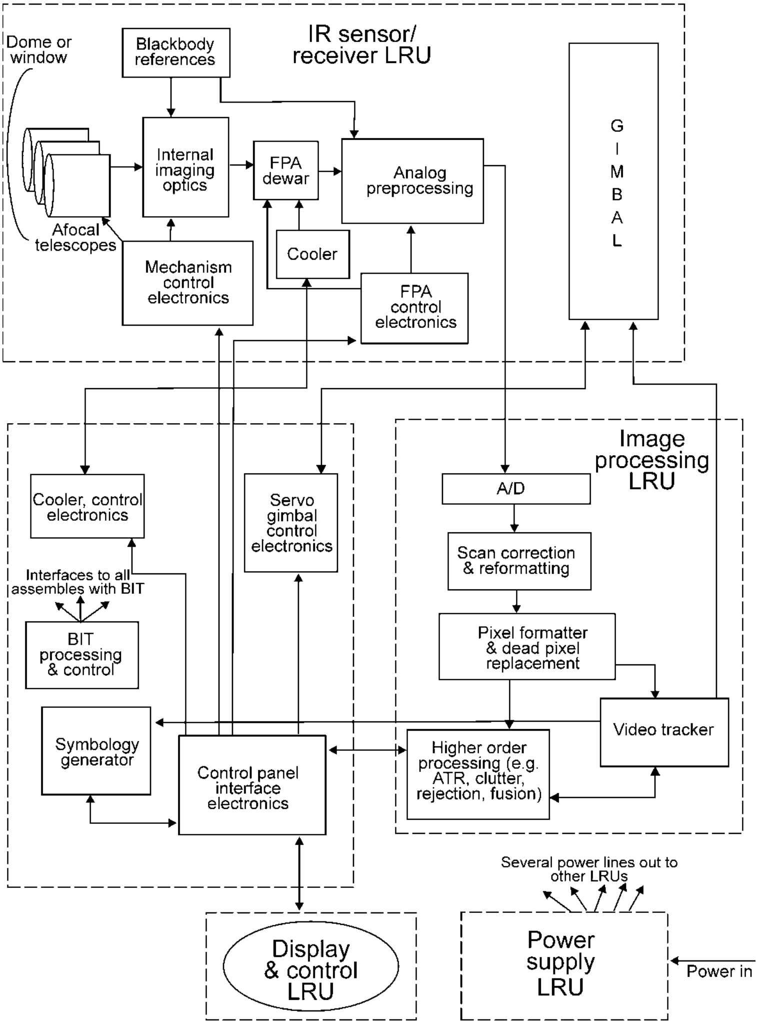

Dome or window
Blackbody references
IR sensor/ receiver LRU
Internal imaging optics
FPA dewar
Analog preprocessing
G I M B A L
Afocal telescopes
Cooler
Mechanism control electronics
FPA control electronics
Cooler, control electronics
Servo gimbal control electronics
Image processing LRU
A/D
Scan correction & reformatting
Interfaces to all assembles with BIT
BIT processing & control
Pixel formatter & dead pixel replacement
Symbology generator
Video tracker
Control panel interface electronics
Higher order processing (e.g. ATR, clutter, rejection, fusion)
Several power lines out to other LRUs
Display & control LRU
Power supply LRU
Power in

Figure B8.22. Representative FLIR system architecture. Reproduced from [7].

spatial frequency are expressed in angular units. By convention, FOV is expressed in degrees, resolution in milliradians, spatial frequency in cycles per milliradian and noise in units of temperatures.

Worldwide, there are over 100 different FLIR systems in operation. The most important of them are described in the literature [7, 21]. Several FLIRs integrate a laser ranger or target designator.

Figure B8.23 shows Falcon Eye, a representative conformal FLIR. This IR device receiver unit features a special gimbal set that allows the 5 in ball to rotate side to side and up and down. This head-steered FLIR adds an excellent degree of tactical flexibility and night situation awareness by allowing the pilot to look in any direction—including directly above the aircraft.

Recent outgrowths of military FLIRs are the infrared search and track (IRST) systems. They are a subset or class of passive systems whose objective is to reliably detect, locate and continuously track IR-emitting objects and targets in the presence of background radiation and other disturbances. They are used in a radar-like manner (usually with a radar-like display) to detect and track objects. Most of

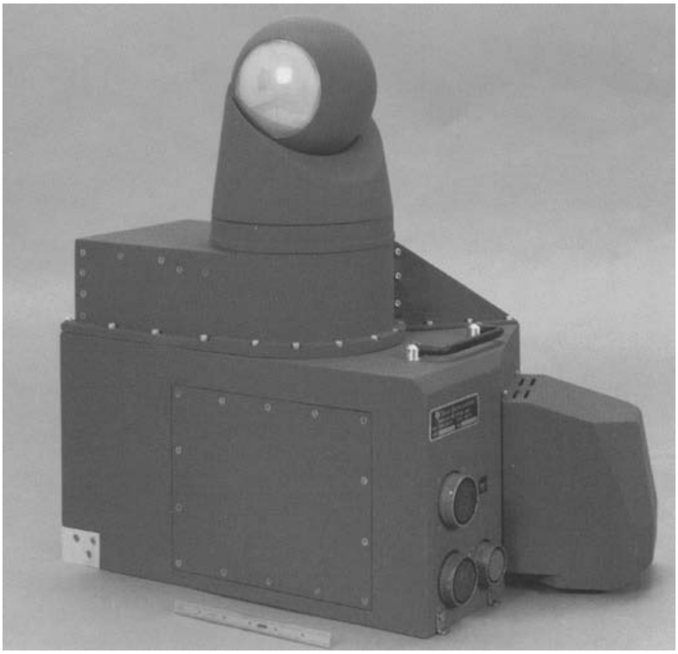


Figure B8.23. Falcon Eye, a representative conformal FLIR (photography courtesy of Texas Instruments).

the current research in IRST systems is concentrated in signal processing to extract target tracks from severe clutter.

Another group of outgrowths of military thermal imagers are airborne line scanners. These are one-dimensional scanning systems that enable creation of a two-dimensional thermal image of the observed scenery only when the system is moving. In contrast to typical thermal imagers with FOV not higher than about 40°, the airborne line scanners can provide FOV up to about 180°. Due to wide FOV airborne thermal scanners are widely used in military aerial reconnaissance.

B8.3.4 Space-based systems

The formation of NASA in 1958, and development of the early planetary exploration programme, was primarily responsible for the development of the modern optical remote sensing systems, as we know them today. During the 1960s optical mechanical scanner systems became available that made possible acquisition of image data outside the limited spectral region of the visible and NIR available with film. ‘Eye in The Sky’ was the first successfully flown long-wavelength sensor launched in 1967. A major milestone was the development of the Landsat Multispectral Scanner because it provided the first multispectral synoptic in digital form. The period followed the launch of Landsat-1 in 1972 stimulated the development of a new series of air- and spaceborne sensors. Since that time, hundreds of space-based sensors have been orbited.

The main advantages of space IR sensors are as follows [7]:

- the ability to tune the orbit to cover a ground swath in optimal spatial or temporal way;
- a lack of atmospheric effects on observation;

- global coverage;
- the ability to engage in legal clandestine operations.

Hitherto, anti-satellite weapons do not exist, so satellites are relatively safe from attack. The disadvantages of satellite systems are protracted and excessive costs of fabrication, launch and maintenance of satellites. Moreover, such operations as repair and upgrade are difficult, expensive and usually not possible.

The space-based systems installed on space platforms usually perform one of the following functions: military/intelligence gathering, astronomy, Earth environmental/resources sensing or weather monitoring. So these functions can be classified as forms of Earth remote sensing and astronomy.

Figure B8.24 shows representative space sensor architecture. It should be stressed, however, that many individual space sensors do not have this exact architecture.

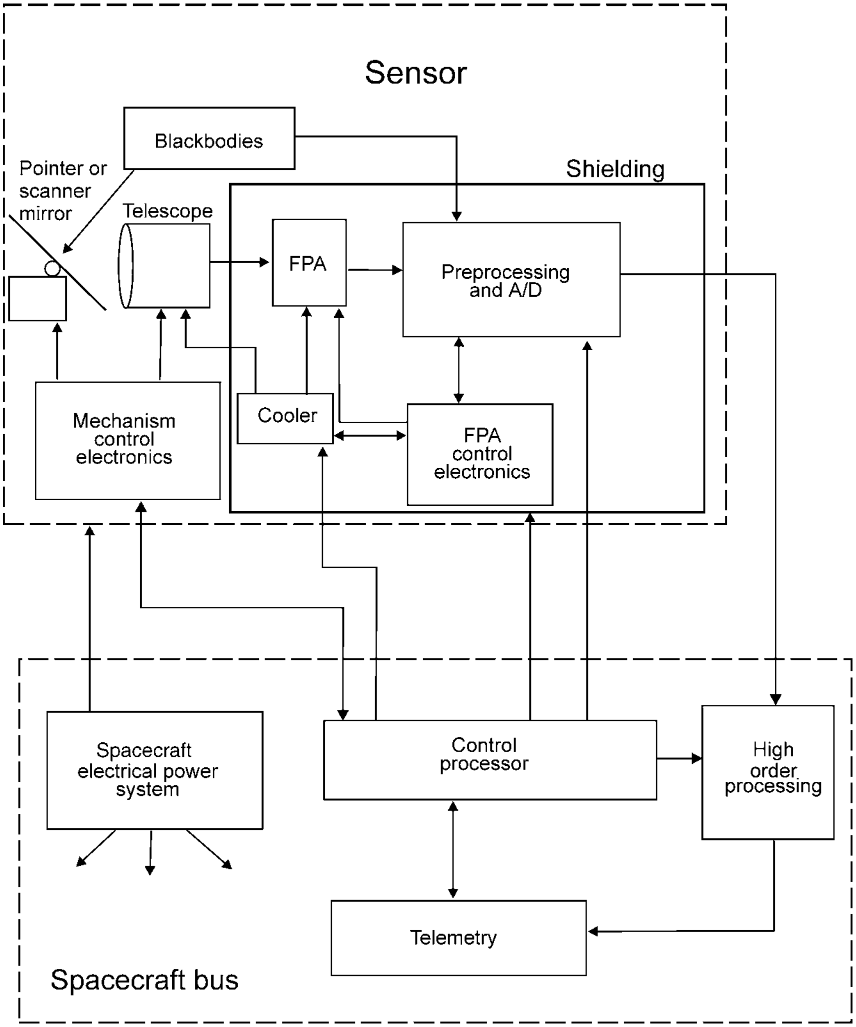

Sensor
Blackbodies
Pointer or scanner mirror
Telescope
Shielding
FPA
Preprocessing and A/D
Cooler
FPA control electronics
Mechanism control electronics
Spacecraft electrical power system
Control processor
High order processing
Telemetry
Spacecraft bus

Figure B8.24. Representative space sensor architecture. Reproduced from [7].

Intelligence and military services from wealthy nations have long employed space-based sensors to acquire information. A satellite-borne IR warning receiver, designed to detect intercontinental ballistic missiles, is a strategic system that protect a large area, or nation. The US spends about $5 billion per year on space reconnaissance [7]. Although the cold war is over, the long-term strategic monitoring to access military and economic might is still important. Intelligence gathering of crop data and weather trends from space has also been used by hunger relief organizations to more effectively forecast droughts and famines. The military also has space-based surveillance for missile launches and additionally, space basing provides excellent viewing geometries for global events as nuclear explosions and environmental changes that the military is concerned about.

Imaging with IR FPAs provides increasingly detailed and quantitative information about relatively cool objects in the space of our galaxy and beyond. Dwarf stars, for example, or giant Jupiter-like planets in other distant solar systems, do not emit much visible and ultraviolet light, so they are extremely faint at these wavelengths. Also, the longer IR wavelengths can penetrate dusty and optically opaque nebulous molecular clouds in interstellar space where new stars and planetary systems are forming.

There are several unique reasons for conducting astronomy in space [7]:

- to eliminate the influence of absorption, emission and scattering of IR radiation;
- to answer basic cosmological and astronomical questions (e.g. formation of stars, protoplanetary discs, extra-solar planets, brown dwarfs, dust and interstellar media, protogalaxies, the cosmic distance scale and ultra-luminous galaxies);
- to observe the Earth's environment (detecting the subtle changes indicating environmental stresses and trends).

B8.4 Noncontact thermometers

Infrared thermometers always measure temperature indirectly in two stages:

- measurements of radiation power in one or more spectral bands;
- determination of an object temperature on the basis of the measured radiometric signals.

Even simple IR thermometers usually consist of five or more blocks (see figure B8.25). An optical objective is usually used to increase the amount of radiation emitted by the tested object and to limit thermometer FOV. The signal at the output of the detector is typically amplified, converted into a more convenient electronic form and finally digitized. A separate visualization block is typically used for presentation of the measurement results.

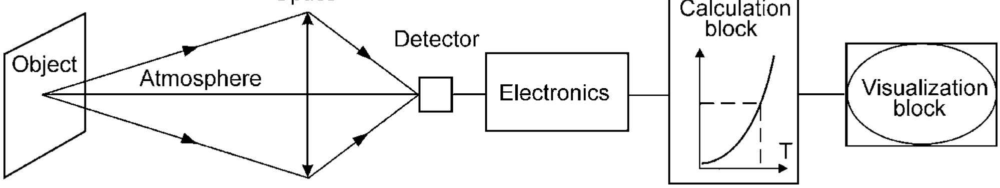


Figure B8.25. General diagram of a simple noncontact thermometer (after reference [22]).

The IR thermometers can be divided into a few groups according to different criteria: presence of an additional co-operating source, number of system spectral bands, number of measurement points, width of system spectral bands and transmission media.

In a passive system, the object temperature is measured knowing the radiation power emitted by the object in one or more spectral bands. With an active system we can get some information about the emissive properties of the tested object by using an additional co-operating source that emits radiation directed to the tested object and measuring the reflected radiation. They are active systems. Active thermometers are more sophisticated, more expensive and so far only in a few applications can they really offer better accuracy than passive systems. Therefore, nowadays, almost all practical noncontact thermometers are passive ones.

In the passive single-band systems, the object temperature is determined using a system calibration chart derived from radiometric calculation of the output signal as a function of blackbody temperature. The temperature of nonblackbody objects can be corrected if only their emissivity over the spectral band is known.

The ratio of the power emitted by a greybody at two different wavelengths does not depend on the object emissivity but only on the object temperature. In passive dual-band systems, the object temperature is usually determined using a calibration chart that represents a ratio of the emitted power in two bands as a function of the object temperature.

At present, at least 95% of systems available commercially on the market are passive single-band systems; passive dual-band systems are rather rarely used; passive multiband systems are still at a laboratory stage of development.

According to the number and location of measurement points, the infrared thermometers can be divided into pyrometers, line scanners and thermal cameras. Pyrometers enable temperature measurement of only a single point or rather of a single sector (usually a circle or a square) of the surface of the tested object. Line scanners enable temperature measurement of many points located along a line. Thermal cameras enable temperature measurement of thousands of points located within a rectangle, square or circle, and create a 2D image of temperature distribution on this area.

Most commercially available noncontact thermometers are pyrometers (see figure B8.26). They are small, light and low-cost systems that have found numerous applications in industry, science, etc. Line scanners are especially suitable for temperature measurement of moving objects and have found applications in the automotive industry, welding, robotics, etc. Thermal cameras offer the greatest capabilities of all discussed types of noncontact thermometer. In spite of their high price, they have found numerous applications such as control of electrical supply lines, heat supply lines, civil engineering, environmental protection, nondestructive testing and many others.

The fixed, inflexible configuration of a noncontact thermometer (like that presented in figure B8.26) is not a good solution in situations when direct sighting due to obstructions is impossible, significant interference is present and electronics must be placed at a safe distance, or very high temperatures exist. In such situations, it is better to use flexible fibre thermometers without the optics block.

B8.5 Radiometers

In general, the IR thermometers discussed in section B8.4 can be treated as a class of radiometers because they determine temperature on the basis of the signal generated by the radiant flux coming to the detector. However, the IR thermometers are designed to measure only temperature and it is usually not possible to use them to measure radiant flux. In our definition, a radiometer is an instrument designed to measure quantities of IR radiation, radiant properties of materials or IR detector parameters.

The IR radiometers can be divided into a few groups according to different criteria: measured quantity, number of spectral bands and number of measurement points. Radiometers enable

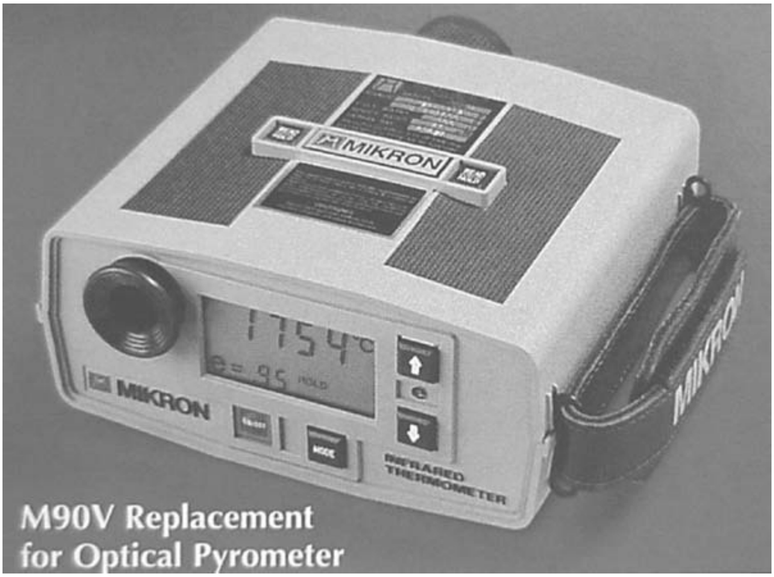


Figure B8.26. M90V pyrometer manufactured by Mikron Instrument Company, Inc. (photography courtesy of Mikron Inc.)

measurement of such quantities of IR radiation as radiant power, radiant energy, radiant intensity, radiance, irradiance, radiant exposure; radiant properties of materials (like emissivity, reflectance and transmittance); parameters of IR detectors (like responsivity and detectivity). However, all these features are possessed by only a small group of radiometers that are generally bulky, laboratory type and expensive systems. Their extreme versatility is usually achieved by a modular approach coupled with an extensive selection of accessories and powerful application software packages which enables the user to tailor a turn-key system to their exact requirements as well as insuring expandability in the future. On the other hand there are radiometers of design optimized for measurements of only a single quantity. Optical power meters are the prime example of radiometers from the latter group (see figure B8.27).

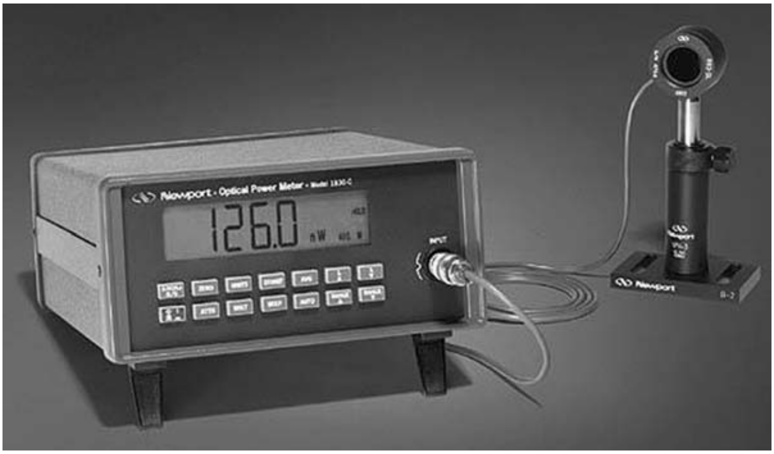


Figure B8.27. Optical power meter model 1830 (see www.newport.com) (with permission from Newport Inc.).

According to the criterion of number of spectral bands, the IR radiometers can be divided into the following groups: single-band radiometers, dual-band radiometers and multiband radiometers. Multiband radiometers enable measurements of one of the above-mentioned radiometric quantities in at least three separate spectral bands. When the spectral bands are narrow and their numbers are high enough, then the multiband radiometers are termed spectroradiometers. In contrast to the situation in noncontact thermometry where the single-band IR thermometers dominate on the market, the spectroradiometers have found a wide area of applications and are the most popular group of the radiometers.

The key component of any spectroradiometer is a module that can be termed the spectral band selector. Its task is to select the desired spectral band from the incoming radiation. This task is achieved by the use of three methods: variable filters, monochromators and Fourier transform (FT) interferometers.

The transmission wavelength of circular (linear) variable filters (VFs) changes continuously (discretely) with the position of the fraction of the filter. Simplicity of design is a great advantage of this type of spectroradiometer; it enables design of small-size, reliable, high-speed and mobile systems. However, using variable filters, it is not possible to achieve very good spectral resolution (typically about 2% of the wavelength). Next, because the system must measure output radiation selected by a variable filter of narrow spectral band and low transmission coefficient, it is necessary to use cooled IR detectors. Cooled sandwich InSb/HgCdTe enabling measurement in the spectral range 2.5–15 μm is a typical option for this type of spectroradiometer.

A monochromator is an optical instrument that uses a dispersing component (a grating or a prism) and transmits to the exit slit (optionally directly to the detector) only a selected fraction of the radiation incoming to the entrance slit. The centre wavelength of the transmitted spectral band can be changed within the instrument spectral region by rotation of the dispersing element. Dispersing prisms, or more often gratings, are used as the dispersing elements in monochromators.

The Michelson interferometer is the spectral band selector in FT spectroradiometers. The interferometer is usually built as an optical instrument consisting of a beam splitter and two flat mirrors arranged so as to recombine the two separated beams back on the same spot at the beamsplitter. One of the mirrors moves linearly in order to produce variable optical interference.

The Michelson interferometer can also be seen as a modulator. From a constant spectral radiation input, a temporal modulation occurs at the detector having a unique modulation frequency for each wavelength of radiation. The modulation frequency can be scaled via the velocity of the mirror movement. This modulated signal registered by the detector is called the interferogram. It is digitized at the rate of at least twice the maximum modulation frequency and a mathematical operation, the Fourier transform, is applied to retrieve the spectral distribution of the input radiation. A calibration with a known source is required in order to obtain quantitative radiometric results.

FT spectroradiometers are characterized by very good spectral resolution and very good sensitivity, better than offered by other types of spectroradiometer. Very good spectral resolution is the effect of use of the interferometer as a spectral selector. Very good sensitivity originates from the fact that the detector is irradiated not only by the radiation from a desired narrow spectral band but also by a full spectrum of radiation coming to the interferometer input. This feature enables design of high-speed, high-spectral-resolution FT spectroradiometers using noncooled or thermoelectrically cooled detectors (typically HgCdTe detectors) instead of bulky liquid-nitrogen-cooled detectors needed in the variable filter or dispersive spectroradiometers. However, the performance of the FT spectroradiometers can be severely reduced even by a very small nonalignment of the optical system which makes this type of spectroradiometer inherently sensitive to shocks and vibrations. Therefore, FT spectroradiometers were, for the last few decades, considered as rather laboratory type equipment that cannot be used in field applications. However, at present, this opinion is outdated as fully mobile FT spectroradiometers are on the market (see figure B8.28).

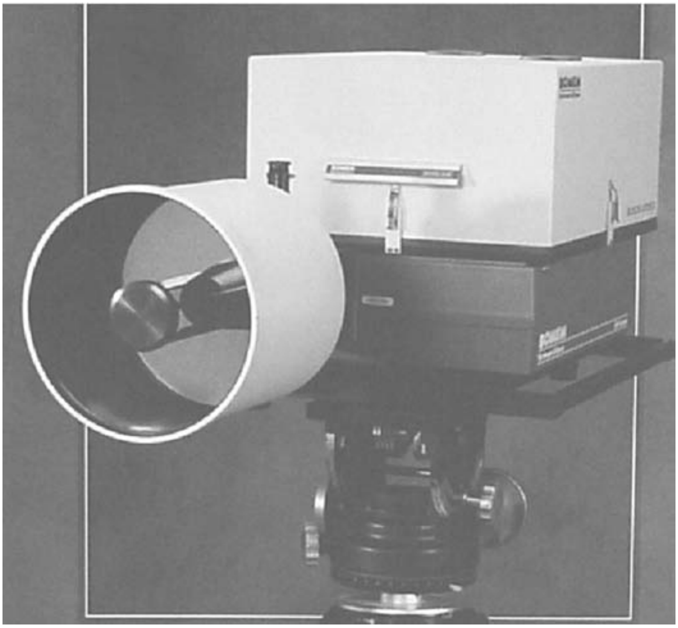


Figure B8.28. FT spectroradiometer (with permission from Bomem Inc.).

The great majority of commercially available spectroradiometers are systems enabling measurement of the spectral distribution of radiation emitted or reflected by a single spot and these systems can be termed the spot radiometers. There also exists another group termed the imaging spectroradiometers because these systems offer some imaging capabilities.

The optical system of any monochromator creates, at the output plane, a series of adjacent images of the input slit corresponding to wavelength. At one time only, one of these images fits to the exit slit and this radiation is measured by the detector located behind the stationary exit slit. Rotation of the dispersing element causes movement of these series images and the radiation from each of them can be measured. If we put an array detector at the output plane of the monochromator optical system, instead of the traditional configuration (an exit slit and a single detector behind), then this series of adjacent images of the input slit corresponding to wavelength will be focused on different parts of the array detector. Therefore, by use of an array detector, we could expect the possibility of simultaneous measurement of the radiation spectrum of different spots within the input slit. However, due to significant aberrations (curved output field, astigmatism) of the optical systems, this cannot be achieved in standard nonimaging spectroradiometers. The aberrations create a situation when only one image from the series—the image that fits the exit slit—is horizontally sharp. By use of a modified optical system with corrected curved output field and astigmatism, we can get sharp images of the input slit corresponding to wavelength focused on different parts of the array detector which enables simultaneous measurement of the spectrum.

The second class of imaging spectroradiometers is multiband (in the literature multispectral or hyperspectral) imaging systems generating simultaneously 2D images of the observed scenery in a number of spectral bands, where this number can vary from a few bands to over a hundred.

A high number of spectral bands being simultaneously recorded is typically achieved by use of a number of dichroic beam splitters, gratings, and linear detectors of different spectral sensitivity regions. The beam splitters separate the incoming parallel polychromatic beam into a few beams of separate

Table B8.4. Basic parameters of the GER EPS-H imaging spectroradiometers (see www.ger.com/epsh.html) (with permission from the GER Inc.)

	Parameters	
Scanner	Kennedy large-size scan head	
Spectrometers	1 Si (VIS/NIR)	76 channels; 0.43–1.05 μm
	1 InGaAs (SWIR 1)	32 channels; 1.5–1.8 μm
	1 InSb (SWIR 2)	32 channels; 2.0–2.5 μm
	1 HgCdTe (LWIR)	12 channels; 8–12.5 μm
	1 InSb (MWIR)	3–5 μm (option)
IR detector cooling	Liquid nitrogen (closed cycle—option)	
IFOV	Choice of 1.25 mrad, 2.5 mrad, 3.3 mrad, or 5.0 mrad	
Swath angle	Up to 90°	
Scan speed	Up to 25 Hz with all bands, continuously selectable	
Pixels per line	512	

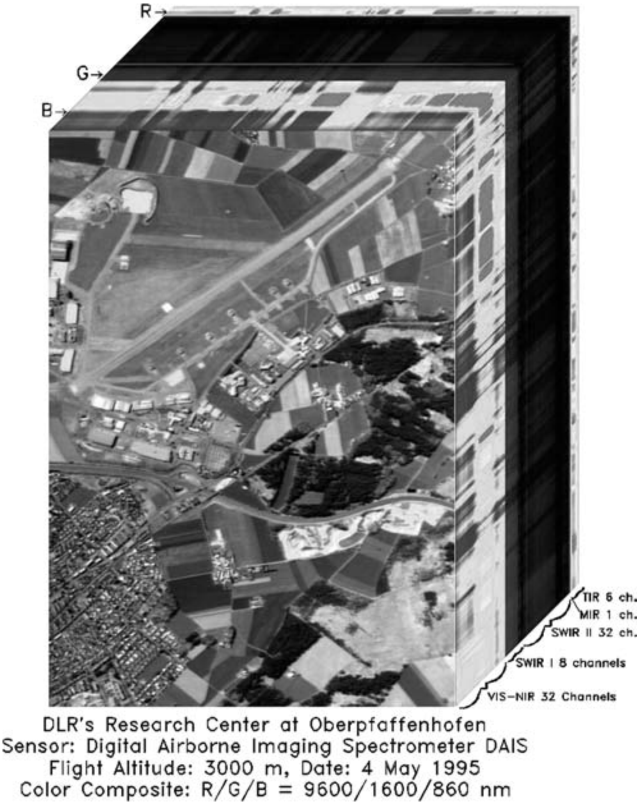

R→
G→
B→
TIR 6 ch.
MIR 1 ch.
SWIR II 32 ch.
SWIR I 8 channels
VIS–NIR 32 Channels
DLR's Research Center at Oberpfaffenhofen
Sensor: Digital Airborne Imaging Spectrometer DAIS
Flight Altitude: 3000 m, Date: 4 May 1995
Color Composite: R/G/B = 9600/1600/860 nm

Figure B8.29. Colour composite image recorded using the Digital Airborne Imaging Spectroradiometer (DAIS) (with permission from German Aerospace Center DLR) (see http://www.op.dlr.de/dais/dais-gal.htm).

spectral bands, for example, see table B8.4. The grating separates the beams further and finally all these spectrally separated beams are focused by output optical objectives at different elements of the linear detectors.

The above-described system enables measurement of flux from a single spot in a number of spectral bands. However, because these systems employ a scanning system (typically the Kennedy-type reflective scanner) and are used in airborne applications, the imaging spectroradiometers generate a two-dimensional image of the land below the aircraft in different spectral bands (see figure B8.29).

B8.6 Light detection and ranging (LIDAR)

LIDAR is an acronym for light detection and ranging, a technique that uses laser light pulses to detect contaminations or aerosol in much the same way that sonar uses sound pulses, or radar uses radio waves. In radar, radio waves are transmitted into the atmosphere, which scatters some of the power back to the radar's receiver. A LIDAR also transmits and receives electromagnetic radiation, but at higher frequency. LIDARs operate in the ultraviolet, visible and IR region of the electromagnetic spectrum.

Different types of physical process in the atmosphere are related to different types of light scattering. Choosing different types of scattering process allows atmospheric composition, temperature and wind to be measured. The scattering is essentially caused by Rayleigh scattering on nitrogen and oxygen molecules, and Mie scattering on aerosols (dusts, water droplets, etc). At low attitudes, Mie scattering is predominant because of the higher cross-section and the high aerosol concentration.

In the LIDAR approach, a laser radiation is transmitted into the atmosphere and backscattered radiation is detected as a function of time by optical receiver. The return time of the reflected or scattered pulses provides range information. In a LIDAR arrangement, the backscattered light is collected by a telescope, usually placed coaxially with the laser emitter. The signal is then focused on a photodetector through a spectral filter, adapted to the laser wavelength. Different kinds of laser are used depending on the power and wavelength required. The lasers may be both cw or pulsed. LIDARs typically use extremely sensitive detectors, which convert the individual quanta of light first into electric currents and then into digital photocounts, which next can be stored and processed on a computer.

In general, a signal is produced by either direct absorption, fluorescence or Raman scattering. Absorption techniques are most straightforward and widely applied. In the atmosphere, for example, long-path absorption spectrometry is used in two wavelength bands—the infrared, where many molecules have characteristic fingerprints, and the ultraviolet (UV) to visible range.

In a typical case, the laser is alternatively tuned to a wavelength within the absorption band of interest (at λ_{on}) and then to a wavelength with negligible absorption (at λ_{off}), so that difference in the signal returned either from a surface or from air- or water-borne particles is recorded. By dividing the two LIDAR signals by each other, most troublesome and unknown parameters are eliminated and the gas concentration as a function of the range along the beam can be evaluated. Such applications require tunable lasers, either tunable diode lasers in the IR or Nd:YAG dye lasers in the UV to visible range.

The principle for different absorption LIDAR (DIAL) is schematically represented in figure B8.30. Let us now assume that wavelength couple (λ_{on}, λ_{off}) is sent simultaneously into the atmosphere. As λ_{on} and λ_{off} have been chosen close enough to exhibit the same scattering properties, the first chimney plume will cause an increase in the backscattering signal, because the concentration of aerosols is larger, but the same increase for both pulses. Conversely, the second chimney plume, which contains a certain quantity of the pollutant, will absorb the backscattered signal at the λ_{on}-wavelength much more strongly than at the λ_{off}-one. From this difference, and using the Beer–Lambert law, one can deduce the specific concentration of the pollutant under investigation versus range. For typical pollutants, such as sulfur dioxide, nitrous oxide, ozone and mercury, detection ranges for the part-per-billion detection level are between 0.5 and 5 km.

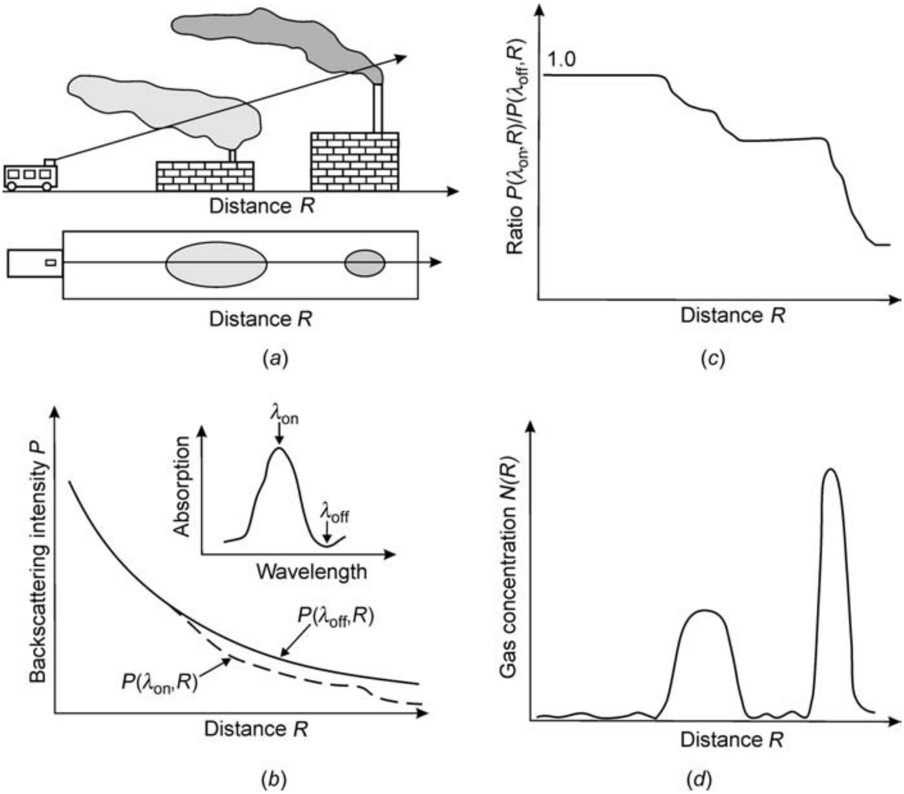


Figure B8.30. Illustration of the principle of different absorption LIDAR (DIAL): (*a*) pollution measurement situation, (*b*) back-scattered laser intensity for the on- and off-resonance wavelengths, (*c*) ratio (DIAL) curve and (*d*) evaluated gas concentration. Reproduced from [23].

The main alternative to direct-absorption spectrometry is Raman scattering. This occurs when photons are inelastically scattered from molecules, exciting them in the process and releasing some photon energy. Thus, Raman return signals are at a different, longer wavelength than the exciting wavelength. Raman cross-sections are much smaller than absorption cross-sections, so the Raman technique works well using high-power lasers only for higher concentrations (hundreds of parts per million) and distances of less than 1 km. Water vapour profiles can be obtained in vertical soundings up to several kilometres in height, and pressure profiles up to tens of kilometres are measurable using Raman signals from atmospheric N_2.

The other major technique, fluorescence spectrometry, has limited use in atmospheric measurement because the return signal is too weak. The technique is, however, an excellent way to monitor solid targets in the biosphere, such as oil, spills, algae bloom patches and forest area. The fluorescent signal from plants originates in the excitation of chlorophyll and other leaf pigments. Fluorescence LIDAR is also a powerful technique for measurements at mesospheric heights where the pressure is low and the fluorescence is not quenched by collisions. This technique has been used extensively to monitor layers of various alkali and alkaline earth atoms (Li, Na, K, Ca and Ca^+) at a height of about 100 km [23, 24].

In addition to monitoring pollutants, LIDAR is widely used to measure wind velocities via Doppler shifts. Recently, improved laser stability has expended LIDAR to more ambitious projects, including the study of winds in the stratosphere.

The main advantage of LIDAR is that it can map the location of chemicals over a wide region. Due to the rapid nature of laser pulses, the time resolution is critical (a few nanoseconds) to get good spatial resolution. Overall, DIAL systems can provide 2D or 3D information of air pollutants. However, most of the existing LIDAR systems have not met the pragmatic deployment requirements of users in industry or government. LIDAR systems are usually complex, large, expensive and require highly skilled personnel for their operation.

B8.7 IR gas sensors

IR gas detection is a well-developed measurement technology. In general, for toxic and combustible gas monitoring applications, IR instruments are among the most user friendly and require the least amount of maintenance.

There are a number of ways by which various IR components can be arranged to produce a gas analyser. The design may be relatively simple, or very complicated depending on the type of analyser for the applications. Figure B8.31 illustrates some of the basic features of an IR analyser.

The basic design is shown in figure B8.31(*a*), which consists of an IR source, bandpass filter and the interaction with the gas sample and detector. A detector is selectively sensitized to the absorption

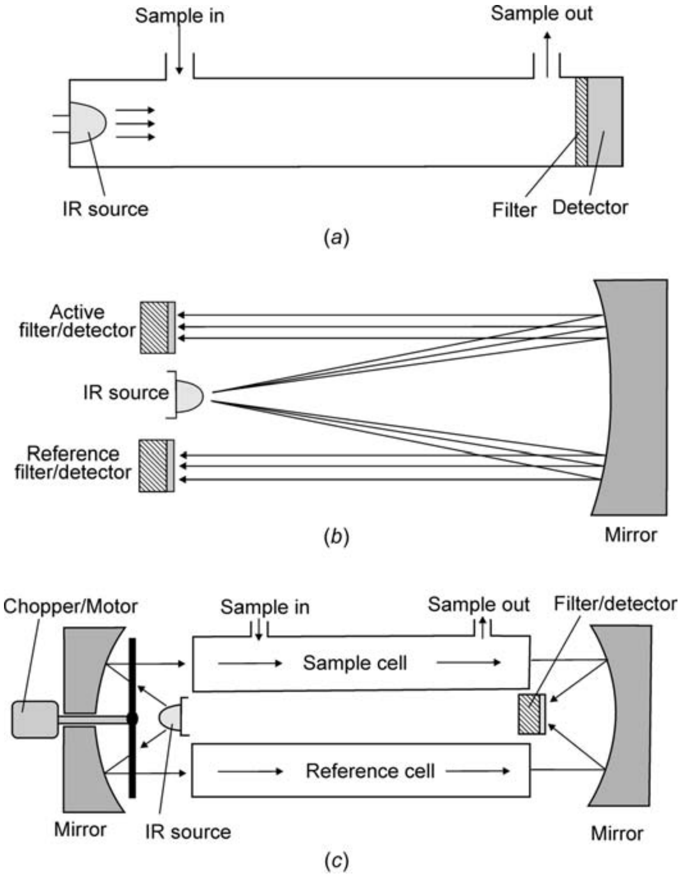

Sample in
Sample out
IR source
Filter
Detector
(a)
Active filter/detector
IR source
Reference filter/detector
Mirror
(b)
Chopper/Motor
Sample in
Sample out
Filter/detector
Sample cell
Reference cell
Mirror
IR source
Mirror
(c)

Figure B8.31. Configurations of the gas analysers: (*a*) a basic gas detector layout, (*b*) a two-detector layout and (*c*) double beams with chopper layout. Reproduced from [25].

wavelength of the gas whose presence is to be detected by the use of a narrow-bandpass optical filter. Clearly then, increased gas concentration in the optical path between source and filtered detector leads to a depression in signal level. The bandpass filter could be placed in front of the light source, instead of placing it in front of the detector.

In practice, in order to reduce false alarms and introduce a level of quantification, it is necessary to provide some calibration. Dependent on the application and instrument manufacturer, this may take the form of a reference chamber containing a known concentration of the gas, or measurement of a reference wavelength just slightly outside the absorption band and/or dual matched detectors.

Figure B8.31(*b*) shows another popular design with layout of two detectors. Modulated flashing IR sources are reflected back to the detectors. In this arrangement, the active detector has a filter for the target gas, while the reference detector has a filter with different wavelength. In such a way, the active detector is used to detect the target gas and the reference detector is used to ignore the target gas. In actual operation, the reference detector provides a base point value (or zero point) while the active detector is used to provide the signal. An advantage of this design is compensation of changes that occur in the detector's sensitivity with time (for example, change in the intensity of the light source).

The design illustrated in figure B8.31(*c*) uses two tubes or cells. One is a reference cell that is filled with a pure target or reference gas, while the other is a sampling cell in which the sample gas passes through. Additionally, a chopper in the form of disc with a number of slots in it is used. As the chopper rotates, it alternately allows the light beam to pass through the sample and reference cells. The detector gets it base reading from the reference cell.

There are many light sources available, ranging from a regular incandescent light bulb to specially designated heating filaments and electronically generated sources. The last sources are used to generate enough radiation at the wavelength of interest for the purpose of detecting the specific target gas. A heated wire filament, similar to that in a pen flashlight, is used in the 1–5 μm spectral range for the detection of most hydrocarbons, carbon dioxide and carbon monoxide. Alternatives include glowbars (rods of silicon carbide) or coils, typically of nichrome alloy resistance wire with high emissivity in the MWIR region.

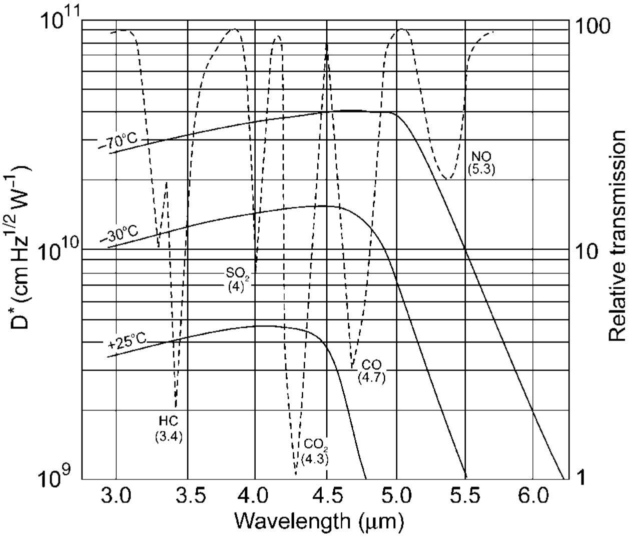


Figure B8.32. Detectivity of lead selenide detectors as a function of wavelength at operating temperatures of +25, −30 and −70°C, along with absorption spectra of some common air polluters.

Figure B8.32 shows the spectral response of lead selenide photoconductors superimposed over the absorption bands of the common air pollutants: HC, SO_2, CO, and NO. Such detectors are optimally matched to the 3–5 μm range. Beyond those wavelengths, the designer tends to be constrained to thermal detectors, the increased cost of the only suitable photon detector (HgCdTe) outweighing the performance benefits in most cases. Hence, the compromise is stated: lower performance over a very broad waveband or higher performance in a more limited one. For most specific applications, the more limited wavebands do contain sufficient structure and thus photon detectors are preferred. On the cost side, thermopiles are the cheapest option, and may be favoured when lower performance is acceptable. In most cases, however, the significant performance benefits of lead salt detectors more than compensate for their relatively small additional cost.

IR detection is applied to numerous applications. The most important are monitoring: in the transport industry (car and truck exhausts), petrochemical industry (gas leaks in refineries and oil rigs) and medicine (carbon dioxide exhalation and anaesthetic gases).

References

[1] Herschel W 1800 Experiments on the refrangibility of the invisible rays of the Sun *Phil. Trans. Roy. Soc. London* **90** 284
[2] Norton P R 1999 Infrared detectors in the next millennium *Proc. SPIE* **3698** 652–665
[3] Ross W 1994 *Introduction to Radiometry and Photometry* (Boston: Artech)
[4] Burnay S G, Williams T L and Jones C H 1988 *Applications of Thermal Imaging* (Bristol: Hilger)
[5] Hudson R D 1969 *Infrared System Engineering* (New York: Wiley)
[6] Rogalski A 2000 *Infrared Detectors* (Amsterdam: Gordon and Breach)
[7] Miller J L 1994 *Principles of Infrared Technology* (New York: Van Nostrand Reinhold)
[8] Walker G 1983 *Cryocoolers* (New York: Plenum)
[9] Couture M E 2001 Challenges in IR optics *Proc. SPIE* **4369** 649–661
[10] Harris D C 1999 *Materials for Infrared Windows and Domes* (Bellingham, WA: SPIE Optical Engineering Press)
[11] Smith W J 2000 *Modern Optical Engineering* (New York: McGraw-Hill)
[12] Lloyd J M 1975 *Thermal Imaging Systems* (New York: Plenum)
[13] Cameron A A 1990 The development of the combiner eyepiece night vision goggle *Proc. SPIE* **1290** 16–19
[14] Csorba I P 1985 *Image Tubes* (Indianapolis: Sams)
[15] Rash C E 1999 *Helmet-Mounted Displays: Design Issues for Rotary-Wing Aircraft* (Bellingham: SPIE Press)
[16] Hewish M 1992 Night-vision goggles *Defense Electron. Comput.* **2** 17–24
[17] Elliott C T, Day D and Wilson B J 1982 An integrating detector for serial scan thermal imaging *Infrared Phys.* **22** 31–42
[18] Kozlowski L J and Kosonocky W F 1995 Infrared detector arrays *Handbook of Optics* Chapter 23, ed M Bass, E W Van Stryland, D R Williams and W L Wolfe (New York: McGraw-Hill)
[19] Kozlowski L J, Vural K, Luo J, Tomasint A, Liu T and Kleinhans W E 1999 Low-noise infrared and visible focal plane arrays *Opto-Electron. Rev.* **4** 259–269
[20] STANAG No. 4349 *Measurement of the Minimum Resolvable Temperature Difference (MRTD) of Thermal Cameras*
[21] Campana S B (ed) 1993 *The Infrared and Electro-Optical Systems Handbook,* vol 5*, Passive Electro-Optical Systems* (Bellingham, WA: SPIE Optical Engineering Press)
[22] Chrzanowski K 2001 *Non-Contact Thermometry—Measurement Errors*, Research and Development Treaties, vol 7 (Warsaw: SPIE Polish Chapter)
[23] Svanberg S 1990 Environmental monitoring using optical techniques *Applied Laser Spectroscopy*, ed W Demtröder and M Inguscio (New York: Plenum) pp 417–434
[24] Wolf J P, Kölsch H J, Rairoux P and Wöste L 1990 Remote detection of atmospheric pollutants using differential absorption lidar techniques *Applied Laser Spectroscopy*, ed W Demtröder and M Inguscio (New York: Plenum) pp 435–467
[25] Chou J 2000 *Hazardous Gas Monitors* (New York: McGraw-Hill)

B9
Organic light emitting devices

Martin Grell

B9.1 Introduction and historic development

In many organic molecules, the absorption of a photon of given wavelength creates an excited state of the molecule (an ‘exciton’) which in turn is capable of re-emitting light of longer wavelength. This is known as fluorescence. The common feature of fluorescent dyes is the presence of alternating single- and double-bonds between carbon atoms (‘conjugated’ units), resulting in delocalized π-electron clouds. A wide range of fluorescent organic dyes, spanning the entire visible and near-IR spectrum, is now available, e.g. for dye laser applications.

Also, the semiconducting properties of a number of organic materials have long been known, and were studied first in crystalline phthalocyanine [1]. Since the 1970s, the amorphous polymeric organic photoconductor poly(vinyl carbazole) (PVK) has been widely studied and is now commonly used for electrophotography (generally, in the form of a charge transfer complex with the electron acceptor 2,5,7-trinitrofluorenone (TNF)). PVK is conceptually a ‘hybrid’ between low-molecular weight and polymeric organic semiconductors. The major difference between a mainchain conjugated polymer such as a polyene or a poly(*para*-phenylene vinylene) (PPV), and a sidechain conjugated polymer such as PVK is that in the mainchain polymer, the π-conjugation extends over more than one repeat unit, leading to a conjugated unit called the ‘effectively conjugated segment’ (ECS) longer than one repeat unit, and with rather different properties.

The capability of semiconducting organic materials to sustain and transport charge carriers (known as electron/hole polarons, or radical anions/cations) opens the possibility to generate excitons by the combination of an electron and a hole polaron rather than the absorption of a photon. The subsequent light emission from an electrically generated exciton is known as electroluminescence (EL). For semiconducting organic crystals, this was first reported by Pope, and Helfrich and Schneider, in the 1960s [2]. In 1983, Partridge reported for the first time on EL from a semiconducting polymer [3]; namely, PVK. Partridge used the common thin film device architecture as shown in figure B9.1. However, OLEDs based on PVK displayed poor brightness and efficiency. This was mainly due to the difficulty to inject electrons into PVK. Ideally, equal number of holes and electrons need to be injected for efficient EL. In 1987, a major breakthrough was reported by Tang and van Slyke of the Kodak group [4]. They introduced OLEDs with a multilayer organic semiconductor architecture. This allowed for better electron/hole balance and lead to devices with much improved brightness and efficiency.

Tang and van Slyke worked with highly fluorescent low-molecular weight organic molecules, and Partridge with a sidechain conjugated polymer. The mainchain conjugated semiconducting polymers that were studied in the 1970s and 1980s such as polyacetylene and polydiacetylene are not fluorescent, because in these polyenes, strong electron–electron interactions break parity alternation between subsequent excited states.

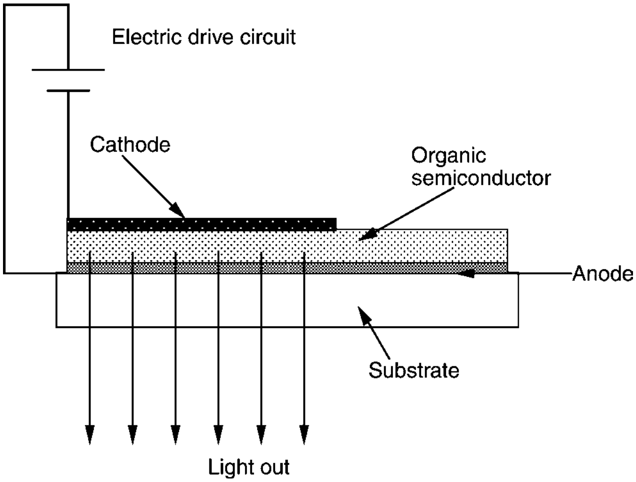


Figure B9.1. The basic OLED architecture. The thickness of the active organic semiconductor film is typically in the order of 100 nm. One of the electrodes (typically, the anode) needs to be transparent to allow for the coupling-out of light.

Thus, in polyenes, the first excited state has no dipole-allowed (fluorescent) transition to the groundstate. In the late 1980s, Friend, Burroughes, Bradley *et al* at Cambridge, experimented with poly(*para*-phenylene vinylene) (PPV), a conjugated polymer containing a phenyl ring in the mainchain. When they tried to establish the dielectric breakdown characteristics of PPV with a view towards its use as gate insulator in organic transistors, they were surprised to find electroluminescence (EL). They published their findings in 1990 [5]. With hindsight, it is surprising that EL from PPV was not discovered earlier, since PPV chemistry and its photoconducting and fluorescent properties had been established previously, much of it due to the work of Hörhold *et al* in Jena [6]. Similarly, another celebrated breakthrough in organic electronics, the discovery of 'synthetic metals' by Heeger, MacDiarmid, Shirakawa *et al* [7], was accidental due to the excessive addition of a catalyst.

The discovery of conjugated mainchain polymer EL has triggered a massive academic and industrial research effort, with the aim of developing and establishing a new flat panel display technology that in principle can replace both cathode ray tube and liquid crystal displays. In recent years, both the development of organic logic circuits based on organic field effect transistors (OFETs), and of organic photovoltaic devices has attracted increasing attention.

B9.2 Common OLED materials and their properties

B9.2.1 Common organic semiconductors

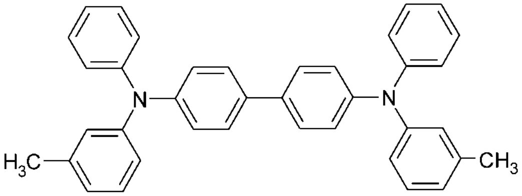
Before discussing OLEDs in detail, it is instructive to have a list of common organic semiconductors that have played an important role in the development of the field. Table B9.1 presents a compilation of a few examples that were selected from the wide variety reported in the literature. Commonly used acronyms are given and some basic properties of the materials are discussed.

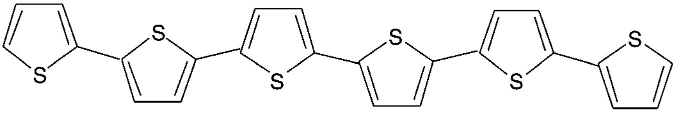
(*a*) *Low-molecular weight materials*

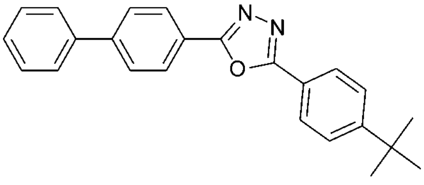
TPD (*N*,*N*′-bis-(*m*-tolyl)-*N*,*N*′diphenyl-1,1-biphenyl-4,4′-diamine) is a hole transporting and weakly luminescent organic semiconductor. 6T (hexithiophene) is one representative of the thiophene family

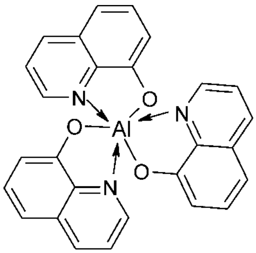
Table B9.1. Selected examples of organic semiconductors and dyes that have played an important role in the development of organic light emitting devices.

(*a*) Low-molecular weight organic semiconductors


TPD


6T


PBD


Alq3

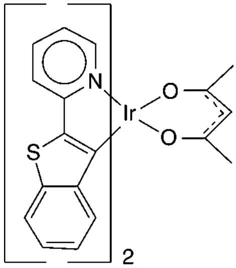
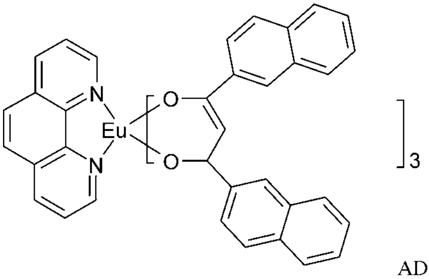
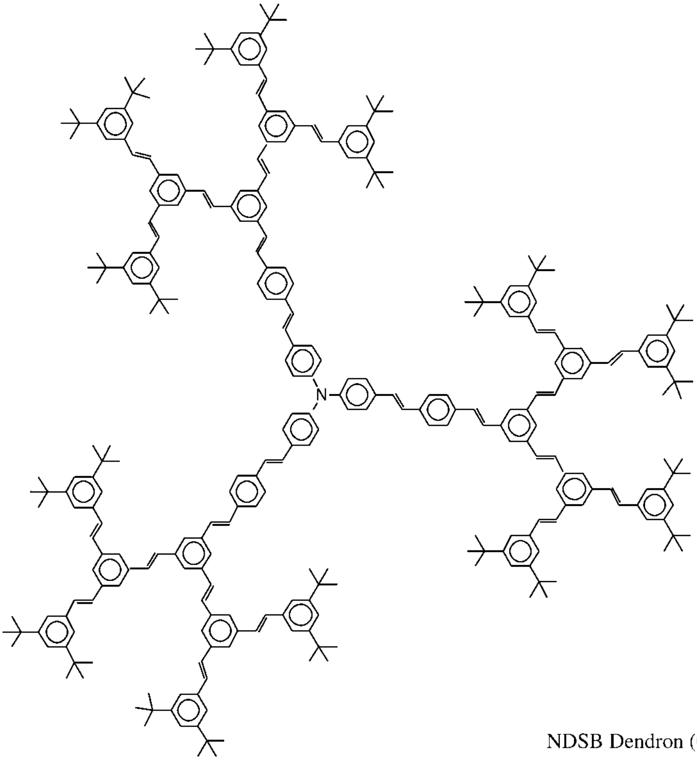

btp$_2$Ir (acac)

Table B9.1. (Continued)

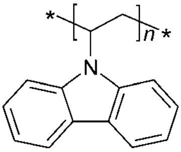


ADS053 RE

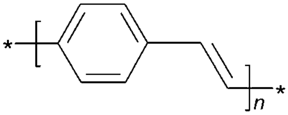


NDSB Dendron (G2)

(*b*) Polymeric organic semiconductors

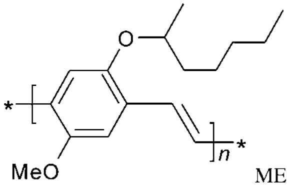


PVK

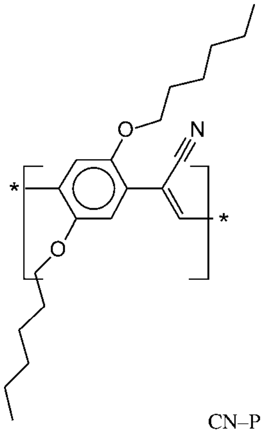


PPV

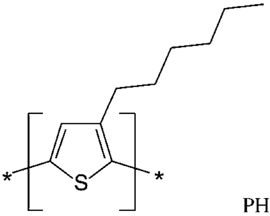
Table B9.1. (Continued)



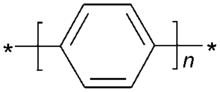
MEH–PPV



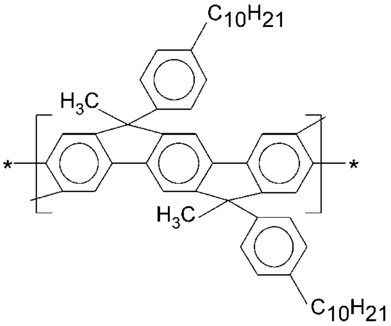
CN–PPV



PHT

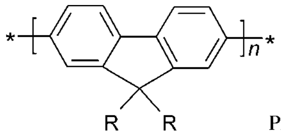


PPP

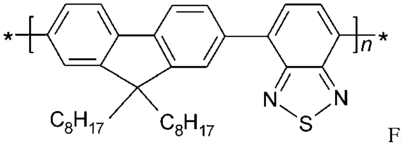

$C_{10}H_{21}$
H_3C
H_3C
$C_{10}H_{21}$

MeLPPP

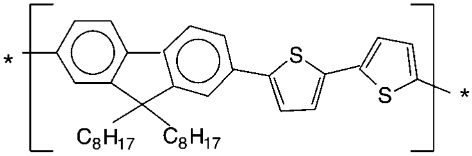
Table B9.1. (Continued)



PF



F8BT



F8T2

of organic semiconductors which are known for their fast hole mobilities. PBD (2-(biphenyl-4-yl)-5-(4-*tert*-butylphenyl)-1,3,4-oxadiazole) is an electron conductor. Both TPD and PBD have been used as carrier injection layers in multilayer device architectures. Alq_3 (tris(8-quinolinolato)aluminum(III)) is an organometallic complex with efficient green electroluminescence and remarkable stability. $btp_2Ir(acac)$ (bis(2-(2′-benzothienyl)-pyridinato-*N*,C-3′)iridiumacetylacetonate) is the representative of a family of phosphorescent dyes that have been used successfully as triplet-harvesting emitters in efficient electrophosphorescent devices. ADS053 RE is the trade name for the red-emitting organolanthanide tris(dinapthoylmethane)mono(phenanthroline)-europium(III). Organolanthanides transfer both singlet- and triplet-excitons to an excited atomic state of the central lanthanide, resulting in very narrow emission lines, i.e. spectrally pure colours. NDSB Dendron (G2) is a nitrogen-cored distyryl benzene second-generation dendrimer. The core displays visible absorption and emission, the *meta*-linked dendronic sidegroups have a bandgap in the UV and for the purposes of charge injection, transport, and light emission can be considered as inert.

(*b*) *Polymeric materials*

PVK (poly(vinyl carbazole)) is a hole transporting and weakly emissive polymeric semiconductor that has been extensively used as photoconductor in photocopiers (in the form of a charge transfer complex with trinitrofluorenone). Conceptually, it is a hybrid between low-molecular weight organic semiconductors and conjugated polymers: The non-conjugated backbone gives it the film forming properties of a polymer, but the semiconducting units are isolated and retain the properties of low-molecular weight carbazole. PPV (poly(*para*-phenylene vinylene)) is a semiconducting, mainchain conjugated and highly emissive polymer. In the form shown in table B9.1, it is insoluble, and it is derived in situ via thermal conversion of a precursor. MEH–PPV (poly[2-methoxy-5-(2′-ethyl-hexyloxy)]-1,4-phenylene vinylene) is a soluble derivative of PPV with somewhat smaller bandgap due to the alkoxy substitution of the benzene ring. Both PPV and MEH–PPV are hole transporting polymers. CN–PPV (poly(2,5-hexyloxy *para*-phenylene) cyanovinylene) is an electron-transporting, low bandgap polymer due to the high electron affinity cyano substitution at the vinylene bond. PHT (poly(hexyl thiophene)) is weakly luminescent, but displays fast hole mobility, in particular, in its regioregular form. PPP (poly(*para*-phenylene)) is a blue emitter, but insoluble in the form shown in table B9.1. Substituting PPP with alkyl sidechains affects solubility, but also twists the rings away from coplanar arrangement. Methylated ladder-type PPP (MeLPPP) enforces coplanar ring arrangement by chemical bonds, and is a blue emitter showing homogeneous broadening only. Poly((9,9-dialkyl)2,7-fluorene)s (PFs) are efficient blue emitters with high carrier mobility. The related copolymer F8BT (poly(9,9-dioctylfluorene-*alt*-benzothiadiazole)) displays a lower bandgap (in the green) and electron transporting properties, as well as partial compatibility with PFs. Consequently, PF/F8BT blends have emerged as alternative to multilayer architectures. F8T2 (poly(9,9-dioctylfluorene-*alt*-dithiophene) is a hole-transporting fluorene copolymer with lower ionization potential than homo-PF.

B9.2.2 Comparing organic with inorganic semiconductors

The molecular nature of organic semiconductors leads to a number of significant physical differences between organic and inorganic semiconductors. The most important are summarized below.

(*a*) *Excitations are localized.* Wavefunction coherence in a conjugated polymer extends over a few (order 5) repeat units (the effectively conjugated segment (ECS) [6]), but not further. Consequently, excitations are localized on the rather short scale of the ECS. The wave vector k is not a good quantum number for a localized excitation. Charged excitations are generally more akin to the concept of the 'radical ion', as familiar from solution-based chemistry, than the concept of a 'polaron' in solid state physics (nevertheless, the term 'hole' is commonly used for radical cations). Neutral excitations or 'excitons' are best described as Frenkel, not Mott-Wannier, excitons.

Localization leads to a strong coupling between excitations and local molecular geometry. In the excited ECS, electron clouds and bond lengths are redistributed. In the case of charged excitations, these geometric relaxations can break the local symmetry, and thus activate vibronic bands in the infrared that are symmetry-forbidden (i.e. Raman- but not IR-active) in the ground state ('IRAVs'). In the case of neutral excitations (excitons), the fluorescence emitted on radiative decay to the ground state may display a relatively large Stokes shift and pronounced vibronic structure.

Typical exciton binding energies, E_b, are in the order 0.2–0.5 eV, and typical exchange energies, i.e. the energetic difference between singlet- and triplet-excitons, are 0.5–1 eV.

(*b*) *Excitations are one-dimensional.* Electronic transition moments are strong and highly directional, parallel to the chain or molecular axis. Some conjugated polymers display liquid crystalline self-organization at elevated temperatures, i.e. the spontaneous parallel alignment of an ensemble of chains [8]. Light polarized parallel to chain alignment then interacts very strongly with the polymer. Also, in aligned polymers, the mobility of charge carriers (radical ions) is enhanced when compared to the non-aligned polymer. Mobility is fastest parallel to the alignment direction, but even perpendicular to the alignment direction, the mobility is faster for an aligned than a non-aligned material [9, 10].

(*c*) *There are no 'dangling bonds'.* Even a very thin organic film always consists of complete molecules, with the chemical coordination at the film surface equal to that in the bulk. This is very different in vapour-deposited inorganic semiconductor films, where atoms at the surface usually are not chemically coordinated in the same way as in the bulk. These 'dangling bonds' distort the band structure at the film surface from its bulk properties. In contrast, surface- and bulk-ionization potentials, and electron affinities of organic semiconductors are generally equal.

B9.2.3 Controlling the bandgap

The relation between optical emission spectra and their perception as colours by human vision is essential for display technology. This shall not be discussed in detail here, a good review is in [11]. The basic facts are that colour perception changes in the order blue–green–yellow–orange–red as wavelength increases from 400 to >700 nm, and that colours are perceived pure when the emission band is narrow. It is therefore essential to control the bandgap of the emissive semiconductors in a display to reproduce the full colour gamut. It is one of the key assets of organic semiconductors that their bandgap can be controlled systematically via modification of their chemical structure. Chemical bandgap tuning can be classified roughly into three types: steric, electronic via sidechain variation and electronic via copolymerization. These are discussed in the following.

The steric control of bandgap

The degree of π-electron delocalization along a polymer backbone is governed by the conformation and configuration of the respective molecule. Conformation can be quantified, e.g. by a dihedral angle φ that describes the relative rotation of a stiff moiety with respect to a neighbouring stiff moiety around a flexible chemical bond connecting the two. The so-called rotational potential $E_{\text{rot}}(\varphi)$ describes the relative energy of the conformation as a function of dihedral angle. Generally, the π-overlap is optimized if successive units are coplanar ($\varphi = 0°$), but disappears if they are orthogonal ($\varphi = 90°$). Hence, in conjugated molecules, there is a contribution to the rotational potential that favours coplanar conformation. However, e.g. in coplanar biphenyl, there would be a clash between the H-atoms in the neighbouring rings for the coplanar arrangement that would be extremely costly in energy. The relative minimum for $E_{\text{rot}}(\varphi)$ is therefore not at $\varphi = 0°$; in the case of biphenyl, $\varphi_{\min} = 23°$ [12]. With the attachment of sidechains, these steric interactions increase, and the twist angle increases from 23° to 45° for *ortho*-dihexyl-substituted polyphenylene. This leads to a ‘blueshift’ (larger bandgap) due to reduced π-conjugation. A prominent example for steric bandgap tuning are the poly(alkyl thiophene)s (PATs), where bandgap tuning throughout the entire visible spectrum by steric effects alone was demonstrated. This is reviewed in [13]; however, PATs are rarely used in OLEDs as their PL quantum efficiencies are very low. Another prominent example for the control of bandgap by more or less bulky sidechains is the case of (non-emissive) poly(diacetylenes) (PDAs), as reviewed in [14].

A method to improve π-overlap, rather than disrupt it by sidechains, is to force rings into a coplanar arrangement by chemical bonds. Let us look at oligomers of *para*-phenylene (PPP), dioctyl fluorene (PFO), and ladder-type *para*-phenylene (MeLPPP), all of which have in common a backbone of *para*-linked benzene rings. In PPP, every ring can twist out-of-plane with respect to its neighbour. In PFO, rings are fused pairwise into a coplanar moiety, there is only one possible rotation per pair, i.e. per two rings. In MeLPPP, all rings are forced into a coplanar arrangement by chemical bonds, there is no degree of conformational freedom in the backbone. Table B9.2 shows the location of absorption maximum (as a measure of bandgap) in dependence of the number of benzene rings for different oligomers (unsubstituted oligo-*para*-phenylene/fluorene-endcapped dihexylfluorene/oligo-ladder-*para*-phenylene).

It is evident that for a given number of rings, the more the backbone is planarized, the smaller is the bandgap.

The electronic control of the bandgap

By introducing either electron-withdrawing or electron-donating chemical groups into a conjugated molecule, the electron affinity and ionization potential are affected, and hence the bondgap changes. Such groups can be introduced in two ways, namely as sidechains or in the mainchain. We discuss

Table B9.2. Absorption maxima for para-phenylene based oligomers versus number of benzene rings (data compiled from [15, 16]).

Number of benzene rings	Oligo-PP (eV)	Oligofluorene (eV)	Oligo-MeLPPP (eV)
3	4.44		3.70
4	4.25		
5	4.15		3.18
6	4.03	3.56	
7			3.00
8		3.43	
9			
10		3.35	

the following examples: Alkoxy-chains attached to PPV rings (MEH–PPV, sidechain modification), CN-groups attached to alkoxy-PPV vinylene bonds (CN–PPV, a case somewhat intermediate between sidechain and mainchain modification), and fluorene copolymers (mainchain modification).

MEH–PPV is an example for alkoxy-substituted PPVs. Its sidechains make MEH–PPV soluble in organic solvents such as THF or chloroform. Sidechains also somewhat isolate chain backbones from each other in the solid film which increases quantum yield over unsubstituted PPV. It was therefore a welcome step forward in the development of conjugated polymers. The alkoxy-linkage of its sidechains to the phenyl backbone ring also changes the electronic structure of the backbone. Alkoxy links have a tendency to donate electrons to the backbone which changes the shape and location of the HOMO. As a result, emission is redshifted compared to PPV, from green to orange.

The case of CN–PPV is somewhat intermediate between sidechain and mainchain modification. In addition to the alkoxy-sidechains in MEH–PPV, highly electron-withdrawing cyano groups are attached to the vinylene bonds. This leaves the conjugated backbone highly electron deficient, thus considerably increasing the electron affinity (by about 0.6 eV [13]). CN–PPVs emit in the red.

Another approach to bandgap control is copolymerization of different conjugated units into the polymer backbone. Copolymers between alkane- and alkoxy-substituted PPV-type polymers are discussed in [13]. Here, we focus on copolymers of fluorene. Polyfluorenes display a ‘blue’ bandgap that is almost indifferent to the type of sidechain attached. Note that sidechains are attached pairwise at the 9-position of the fluorene ring which itself is not part of the conjugated backbone. Consequently, there is little electronic impact of the sidechains on the backbone properties. However, strictly alternating copolymers of fluorenes with comonomers having different electronic properties were prepared, such as F8BT and F8T2. For both F8BT and F8T2, the resulting bandgap is reduced, and they both emit in the green–yellow region. The reduction in the size of bandgap has different reasons: In the case of F8BT, the benzothiadiazole comonomer has a higher electron affinity E_a than fluorene, thus leading to a polymer with higher E_a. In the case of F8T2, the two thiophene groups have a lower ionization potential I_p than fluorene, thus leading to a polymer with lower I_p. Thus, copolymerization not only allows control of bandgap, but also I_p and E_a in a predictable manner, and a large number of fluorene copolymers have been synthesized and studied. For a review, see [17]. Both the polymers have found interesting applications: some of the most efficient organic EL devices have been built from blends of a minority amount of F8BT as electron injecting/transporting material in PFO as hole injecting/transporting material. The miscibility (or at least, slow separation on the spincoating timescale) of PFO and F8BT is highly exceptional, and allows for the preparation of single layer devices with the properties of double layer devices. F8T2, on the other hand, is an excellent material for p-type OFET channels [9].

B9.3 Device preparation and characterization

Practical OLEDs conventionally use a film of indium tin oxide (ITO) on a glass substrate as anode. ITO is a highly doped semiconductor that exhibits almost metallic conductivity ($\sim 20\,\Omega/\square$), and is transparent to allow for the out-coupling of light. Onto the anode, one or several organic layers are prepared. Nowadays, ITO is usually coated with a PEDOT/PSS synthetic metal film, for reasons to be discussed in section B9.4.2. For a review of the remarkable properties of PEDOT/PSS, see [18]. Then, one or several organic semiconducting layers will be deposited; their functions are discussed in section B9.4.

Obviously, it is tempting to replace ITO completely by flexible sheets of PEDOT/PSS prepared on a plastic substrate which are now commercially available. For some small, low-resolution displays this is already possible. However, the most conductive PEDOT/PSS sheets to date still display a sheet resistance about one order of magnitude larger than ITO. This leads to a drop in the applied voltage across a large display and compromises the addressing of pixels far away from the voltage source.

The preparation of polymeric and low-molecular weight organic semiconductor layers is typically very different, and gives rise to two separate ‘cultures’ of organic semiconductor research. Polymers are typically processed from solution. Typical processes are spincasting, which is the conventional laboratory technique, or ink jet printing, which is of increasing interest for industrial production. Low-molecular weight molecules are typically processed by evaporation.

B9.3.1 Solution processing

A key asset of organic semiconducting polymers is that they can be molecularly engineered to be soluble. This allows for the preparation of good quality, uniform thin films over large areas by the spincoating (or spincasting) technique. Much of the interest and momentum in semiconducting polymer research results from this cheap and easy technique to make quality films of arbitrary size.

A typical spincasting solution has a concentration of 5–20 mg of polymer per millilitre of an organic solvent, such as toluene, xylene, tetrahydrofurane (THF), or chloroform. Before use, all solutions should be filtered through a microporous filter, with pore sizes in the range 0.2–0.45 μm.

The principle of a spincoater is shown in figure B9.2.

A drop of polymer solution is placed onto a substrate that is held down on a turntable by vacuum suction. As the turntable starts rotating, the solution spreads out into a thin film that wets the whole

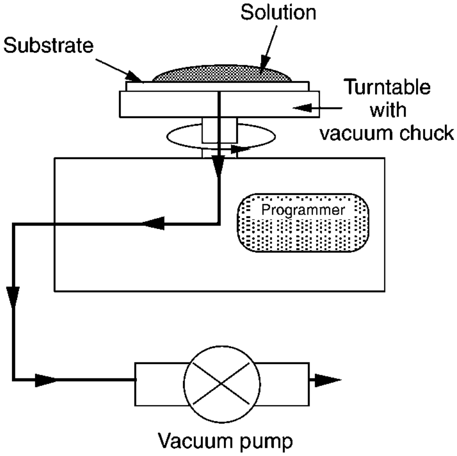


Figure B9.2. A spincoater.

substrate with uniform thickness. Typical 'spin speeds' are in the range 1000–4000 rpm, with spin times in the order of 1 min. Under rotation, the solvent evaporates and leaves behind a thin, uniform polymer film. In principle, there is no limit to the size of the film. In industry, films in the size of TV screens are spincoated routinely by automated systems. The thickness d of the resulting film is controlled by solution concentration c, viscosity η and spin speed ω according to equation (B9.1)

$$d \propto \frac{c\eta(c)}{\sqrt{\omega}}. \tag{B9.1}$$

The proportionality constant depends strongly on solvent, substrate, and other factors. Since η strongly depends on concentration, concentration has a much stronger impact on the thickness of the resulting film than spin speed. d typically will be in the order of 100 nm, with thickness control within a few nanometre via spin speed. Note that the thickness and quality of the resulting films depend on the solvent used. Sometimes, in particular when solvents of low volatility were used, it is advisable to dry films after spinning at moderatly elevated temperature (40–60°C) under vacuum.

Often, multilayer architectures of several layers with different organic semiconductors are required. To make these by spincasting, subsequent layers have to be spun from mutually exclusive solvents (so-called 'orthogonal' solvents), otherwise the first layer dissolves on spinning the second. This can be a challenge to material chemistry. The conjugated polymer community has recently started to use polymer blends deposited from a common solvent in one spin cycle, instead of multilayer architectures.

Drawbacks of the spincoating technique are the need for soluble materials, the sometimes limited options for deposition of multilayer devices, the waste of material that flies off the edge of the substrate on spinning, poor quality of films if the solvent evaporates very fast or does not wet the substrate well, and its limited use for low-molecular weight materials.

B9.3.2 Vapour deposition

For low-molecular weight materials, the method of choice for device fabrication is often vapour deposition instead of spincoating. The apparatus is shown in figure B9.3.

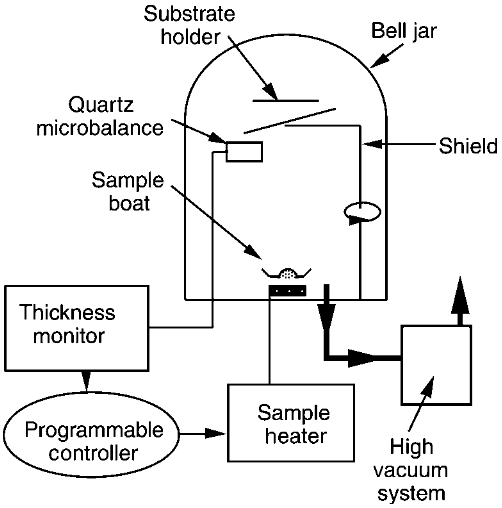


Figure B9.3. An evaporator system for the vapour deposition of small-molecule organic semiconductors, or metals.

In a vapour deposition chamber, high vacuum ($<10^{-6}$ bar) is required. This is usually established with a two-stage pumping system, comprising a rotary pump for a rough pre-vacuum, and an oil diffusion or turbopump for high vacuum. Material is placed into a metal 'boat' and is heated to evaporate and to condense on the substrate. This technique is applicable only to small molecules; polymers do not evaporate. A quartz microbalance is used to monitor the thickness and the deposition rate; this can be fed back via a controller to the sample heater to tune to a programmable deposition rate. Co-evaporation of different materials at the same time is an option, but to arrive at defined blend compositions, careful calibration of the heaters is required. Multilayer architectures, however, are easily made. Tang and van Slyke's first double-layer organic light emitting device was made by vapour deposition [4].

In some cases, exceptional control over the morphology of resulting films can be exercised by evaporation rate, and most importantly, type and temperature of the substrate. Higher temperature allows molecules to reorganize into more ordered structures immediately after deposition, but before a dense film is formed. The resulting films may have different crystal structures, and different orientation of the crystal axis with respect to the substrate; but these may be specific to certain substrates, which are not always useful for device fabrication. Control of morphology via evaporation conditions has been explored in much detail, e.g. for hexithiophenes [19].

Evaporators are also useful for the deposition of metal cathodes, which act as electrical contacts for devices, and can be applied to vapour deposited organic devices without breaking the vacuum. It is a drawback of the generally simpler and cheaper spincasting procedure that the subsequent evaporation of metal cathodes requires an additional high vacuum processing step. Recent polymer OLED research has therefore explored alternative metal deposition procedures that require no vacuum steps [20].

Device characterization

The electrical and optical properties of OLEDs are essential for their applications, and are characterized by a wide variety of methods. The most basic characterization of a conjugated material is by recording its absorption and photoluminescence (PL) spectra, and determining photoluminescence quantum yield (photons out per photons in). The relevant instrumentation and interpretation of such spectra has been discussed widely in the literature [21]. Instead, we discuss here the basic techniques specific to the use of such materials in OLEDs.

Electrical characterization

The current–voltage (C/V) characteristic of an OLED is determined simply by connecting the device electrodes to an electrical source-measure unit. Typically, voltage is ramped up continuously up to the order of 30 V or more, although for an OLED with low onset voltage, up to 10 V or less may be sufficient and higher voltages will lead to a damaging level of current. The typical range of current densities in an OLED under above voltage cycle will be 0.1 (or less) mA cm^{-2} at the light emission 'onset', up to the order of 100 mA cm^{-2} under the highest reasonable (i.e. non-destructive) drive voltages. Current sensitivity for a typical sample device, with typically a few square millimetre active area, should be at least 1 μA.

Often, while undergoing C/V characterization, the luminescent intensity emitted from an OLED is detected simultaneously by a photodiode. Luminescence/voltage characteristics is typically plotted into the same graph as current/voltage, resulting in current/voltage/luminescence ($C/V/L$) characteristics. However, the response of the photodiode is usually not calibrated, and luminescence results are presented in 'arbitrary units' (a.u.). Note that there is no physical current or luminescence 'onset' voltage for an OLED; practically, however, an 'onset' voltage is often reported. This only makes sense if a definition of 'onset' is given.

Quantifying 'brightness'

One of the most interesting properties of an OLED is how 'bright' it is, and how efficiently it converts electrical energy into light. The most straightforward quantity to characterize efficiency is the internal quantum efficiency η_{int}. η_{int} is the ratio of photons generated per two charge carriers injected (two because an exciton is formed from two carriers). To optimize η_{int}, we have to:

– *balance carrier injection.* An excess of one type of carriers means not all injected carriers can find a partner of opposite sign to form an exciton, and will exit the device at the opposite electrode as a 'blind' current. Unpaired carriers contribute to the number of carriers injected, but not the luminescence. Methods to optimize carrier balance are discussed in section B9.4.2.

– *emit as many as possible photons from each generated exciton.* Methods to maximize photon yield are discussed in section B9.5.

The 'external quantum efficiency' η_{ext} is defined as the number of photons coupled out from a device per two carriers injected. Remarkably, η_{ext} is often considerably smaller than η_{int}, often in the order $\eta_{\mathrm{ext}} \approx (1/8)\ \eta_{\mathrm{int}}$. Approaches to optimize η_{ext} are discussed in section B9.5.5.

Both internal and external quantum efficiencies are defined in terms of the fundamental electrical and photophysical properties of OLED operation. Their measurement, however, can be rather intricate. There is no direct approach to η_{int}. To measure η_{ext}, we have to count photons, e.g. in an 'integrating sphere' or with calibrated photometers.

Instead, device performance is often reported in terms of photometric quantities. Photometric quantities are 'physiological' units, i.e. they consider the response of the human eye as well as purely physical quantities such as the power of radiant flux. The most important photometric quantities are the luminous flux F with unit lumen (lm), and the luminous intensity I with unit candela (cd). The luminous intensity I is related to the luminous flux F by $I = \mathrm{d}F/\mathrm{d}\Omega$, with the solid angle Ω in sterad [11]. To give a characteristic that is independent of the arbitrary size of the OLED, the luminous intensity per unit area or 'Luminance', $L = I/A$ in units cd m^{-2}, is used.

The relation between the physiological quantity luminous flux F (in lumen) and the physical quantity radiant flux per unit wavelength P (in W/nm) is given by equation B9.2 [11]

$$F = K \int_{\lambda_1}^{\lambda_2} V(\lambda) P(\lambda)\,\mathrm{d}\lambda \tag{B9.2}$$

where $V(\lambda)$ is the dimensionless 'photopic efficacy' that describes the spectral sensitivity of the human eye, and $P(\lambda)$ is the radiant flux per unit wavelength (in W nm^{-1}). Note that radiant flux is a physical unit and is measured in watts, while luminous flux in lm is the corresponding physiological unit. The interval λ_1 to λ_2 is the wavelength interval wherein $P(\lambda)$ is different from zero, and K is the 'absolute luminous efficiency', $K = 678.8\ \mathrm{lm\,W^{-1}}$. $V(\lambda)$ is non-zero within the range of visible wavelengths ($\approx$ 380 to 750 nm) only, and is normalized to 1 at the wavelength of maximum sensitivity of the human eye, $\lambda_{\max} = 555\ \mathrm{nm}$; thus $V(\lambda) \leq 1$. The above discussion applies to the bright-adapted eye ('photopic vision') which is relevant for display technology. In the dark-adapted state ('scotopic vision') a different K and $V(\lambda)$ apply.

Conveniently, calibrated cameras that give luminous intensity in cd are commercially available. Therefore, device efficiency is often expressed in terms of luminous intensity/electric current through the device (unit cd A^{-1}), rather than in terms of η_{ext}.

Another common quantity to characterize device efficiency is the power efficiency expressed in lumen/watt (lmW^{-1}). The 'watt' in lmW^{-1} here refers to a watt of electric energy driven through the device, i.e. current × drive voltage, not a watt of radiant flux. To optimize power efficiency, we have to optimize not only η_{ext} but also have to minimize the drive voltage required. Section B9.4.1 discusses the method of achieving it.

B9.4 Physics of OLED operation

In fluorescence, excitons are created by the absorption of light, while in EL, excitons are created by electron and hole polaron ‘capture’. Polarons first have to be injected from the electrodes, and migrate towards each other. They then form an exciton that sometimes can decay under the emission of light. The variety of electrical and photophysical processes involved are summarized in the figure B9.4. We discuss the most important of these processes in detail.

B9.4.1 Charge carrier injection

The first step towards exciton formation in an OLED is the injection of holes from the anode/electrons from the cathode under an applied voltage. Good injection is a considerable challenge, in particular, since we need to inject carriers of both signs into the device. Carrier injection is controlled by the workfunction Φ of a metal electrode relative to the electron affinity E_a of the semiconductor for electron injection, and relative to the ionization potential I_p of the semiconductor for hole injection.

A level diagram shown in figure B9.5 is often used to illustrate carrier injection. Note that due to the molecular nature of organic semiconductors, there are no surface ‘dangling bonds’ that can distort bulk energy levels. In figure B9.5, left, there is an ‘injection barrier’ of 0.5 eV for holes from ITO into PPV, and 0.3 or 1.7 eV for electron injection from Ca or Al into PPV, respectively. For electron injection from ITO, there would be a large barrier of 2.2 eV. Thus, the use of electrodes made of unlike metals defines a forward and reverse bias for the OLED.

The right-hand side part of figure B9.5 shows the same device (assuming a Ca cathode) under a forward bias. Carriers can now overcome injection barriers by tunnelling, with tunnelling distances $t_{h/e}$ for holes/electrons, respectively, given by equation (B9.3)

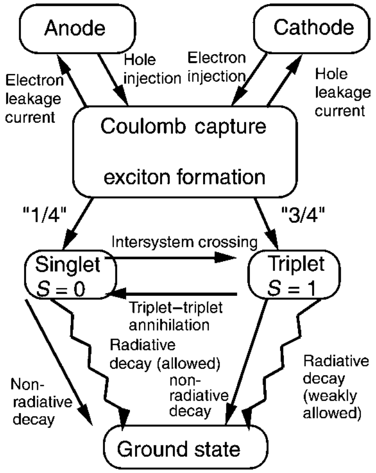

Anode
Cathode
Hole injection
Electron injection
Electron leakage current
Hole leakage current
Coulomb capture
exciton formation
"1/4"
"3/4"
Intersystem crossing
Singlet S = 0
Triplet S = 1
Triplet–triplet annihilation
Radiative decay (allowed)
Non-radiative decay
non-radiative decay
Radiative decay (weakly allowed)
Ground state

Figure B9.4. A chart describing the formation and decay of excitons in organic EL devices. Adapted from [22].

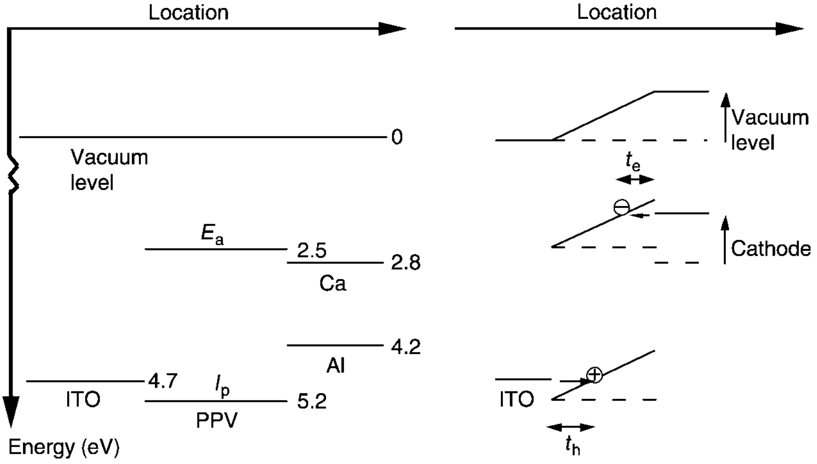


Figure B9.5. Energy levels in a single layer organic EL device with PPV emissive layer. Left: no bias voltage applied. Right: a voltage is applied in forward bias.

$$t_{h/e} = d\frac{\Delta V_{h/e}}{V_{bias}} \tag{B9.3}$$

with d the semiconductor film thickness, ΔV the respective injection barriers, and V_{bias} the applied forward bias. The resulting tunnelling current density/voltage characteristic $j(V_{bias})$ is described by the equation of Fowler and Nordheim ('F–N tunnelling'). For a detailed discussion, see, e.g. [23]; however, it is obvious that injection barriers should be minimized or entirely absent for good carrier injection. A metal–semiconductor junction is termed 'ohmic' if current density across the device is limited by the transport of carriers across the semiconducting film, rather than by the injection barriers at the interface. In that case, carrier transport is termed 'space charge limited conduction' (SCLC). The problem of carrier injection can be considered solved in the case of SCLC. Practically, injection barriers of 0.3 eV mostly lead to ohmic behaviour [24].

Generally, ohmic injection requires high workfunction anodes and low workfunction cathodes. Considerable effort has been devoted to increase the workfunction of the transparent ITO anode by a variety of physicochemical treatment cycles [25]. Recently, it has become common to coat ITO with a thin film of the high workfunction synthetic metal PEDOT/PSS [18] ($\Phi = 5.2\,\text{eV}$). As cathodes, low workfunction materials such as Ca are commonly used. These require protection from ambient atmosphere, otherwise they would rapidly degrade. This can be provided by encapsulation, or by capping with a more stable metal such as Al.

To discriminate experimentally between barrier-type and ohmic injection, it is common to compare $j(V_{bias})$ characteristics of devices of different thickness d. F–N tunnelling depends only on the applied field $E = V_{bias}/d$, thus in a plot of j versus E, all characteristics will coincide regardless of d. In the case of SCLC, j will follow Child's law, $j \sim V_{bias}^2/d^3$. Thus in the plot j versus E, the characteristics will not coincide for different d, but will in the plot jd versus E. The situation is more complex when injection of both electrons and holes may occur. Therefore, the above experimental procedure is to be carried out in symmetric devices, e.g. Au/semiconductor/Au for holes and Ca/semiconductor/Ca for electrons. This will ensure single carrier currents.

B9.4.2 Charge carrier transport

After injection, carriers drift across a device under the pull of the local field, which depends on the applied bias voltage, device thickness, and distribution of space charges in the device. The carrier drift velocity v depends on field E as $v = \mu\mathrm{E}$ with the carrier mobility μ. Experimentally, mobilities can be determined by the time-of-flight (TOF) technique [10], analysis of the $j(V)$ characteristics of space charge limited currents [26], or the analysis of output characteristics of organic field effect transistors [9, 27].

In single crystals of low-molecular weight organic semiconductors, at low temperatures sometimes a band-like (coherent) charge carrier transport is observed with very high charge carrier mobilities in the order of $100\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ [28]. However, this requires elaborate single crystal preparation that is inconsistent with conventional layer deposition onto device substrates such as ITO. In practical situations, one generally finds (incoherent) hopping type carrier transport, which can be described as a directed random walk through a medium characterized by both energetic and positional disorder. This is quantitatively described by a model developed by Bässler in Marburg [23]. Bässler arrives at a rather complex equation to describe both field- and temperature-dependence of carrier mobility μ, which is further complicated in the presence of traps. At constant temperature, the logarithm of mobility is predicted to be proportional to $\sqrt{E}$. As long as $[\mu(E) - \mu(E \to 0)] \ll \mu(E \to 0)$, this can be approximated by equation (B9.4)

$$\mu(E) = \mu(E \to 0) + k\sqrt{E} \tag{B9.4}$$

where k can be both positive or negative, depending on the relative size of disorder parameters describing positional and energetic disorder, respectively. Hence, experimentalists often plot mobility data versus $\sqrt{E}$ [10].

Mobilities in the incoherent transport regime are typically much smaller than the coherent transport, ranging roughly from 1 to $10^{-8}\,\mathrm{cm^2\,V^{-1}s^{-1}}$ for common materials. However, due to the short distance charge carriers have to travel in a typical OLED (film thickness ~100 nm), OLEDs can operate well with rather low charge carrier mobility. For example, the hole mobility in both PPV and dialkoxy-PPV is in the order $10^{-6}\,\mathrm{cm^2\,V^{-1}s^{-1}}$ [26, 29]. Nevertheless, PPV and its derivatives are highly successful OLED polymers. The situation is quite different for organic field effect transistors (OFETs), because carriers have to travel from source to drain across a channel of at least several micrometres.

Practical materials are often characterized by the presence of charge carrier traps. Traps are localized sites with either electron affinity higher than the electron affinity of the bulk material (electron traps), or the ionization potential lower than that of the bulk semiconductor (hole traps). Common trap sites are impurities (e.g. catalyst residues), chemical defects (e.g. sites that have oxidized during OLED operation), and grain boundaries in partially crystalline materials. In the presence of traps, it is not appropriate to describe carrier transport in terms of a single mobility. Instead, we find dispersive charge carrier transport. In a time-of-flight mobility experiment, no clear transient can be assigned in the case of dispersive transport. Instead, transit times vary widely (over several orders of magnitude) due to 'late arrival' of carriers that had been trapped for various lengths of time. Typically, in the case of dispersive transport, the apparent mobility deduced from the time-of-flight experiment is longer when a thick sample is used. This ambiguity signals the breakdown of the assumption of a defined carrier mobility.

Traps are roughly classified as either shallow or deep. Shallow traps are those with trap depths of only a few kT, so that trapped carriers can be thermally reactivated into the conduction band, while deep traps are considerably deeper than kT and will release carriers only when the external bias becomes sufficiently high (in effect, the carrier has to be re-injected back into the transport band against an injection barrier defined by the trap depth). Deep traps are a serious problem, as trapped carriers effectively have zero mobility and thus contribute to space charge but not to the current. The general observation that organic semiconductors do intrinsically carry only one type of carriers (i.e. an organic semiconductor is either hole or electron transporting, but only very few are 'ambipolar') is thought to be the result of carrier-specific deep traps. To avoid deep traps, materials have to be as pure as possible, the formation of defects during operation has to be minimized by the exclusion of oxygen and water (i.e. device encapsulation), and some materials have to be prepared as amorphous ('glassy') films rather than semicrystalline, because crystallite grain boundaries represent deep trap sites. These requirements define some of the extraordinary challenges on organic semiconductor chemistry.

Colloquially, the terms hole (electron) transporting layer (HTL/ETL) are used to denote a material that at the same time facilitates low-barrier or barrier free hole (electron) injection and hole (electron) transport without deep traps. Strictly, injection and trap-free transport are different phenomena; however, they are both closely related to the location of I_p or E_a, respectively. For example, in the case of an ETL, electron injection will be facilitated by high E_a. At the same time, an impurity will act as electron trap only if the E_a of impurities is higher than that of the host material. Thus, high host E_a makes it less likely that an impurity can act as electron trap.

B9.4.3 Electrical device optimization

Assuming that carrier traps can be avoided, it is fair to say that the minimization of injection barriers is much more important for the optimization of OLED efficiency than charge carrier mobility. However, it is generally difficult to achieve ohmic (i.e. barrier free) injection at both electrodes for a given organic semiconductor. The breakthrough towards efficient organic EL devices comes from a simple but ingenious idea. Tang and van Slyke from the Kodak group have manufactured a bilayer device consisting of a low-ionization potential, hole transporting diamine layer and a high electron affinity, electron transporting Alq_3 layer, which is also an efficient green emitter [4]. They also used extremely thin layers of organic semiconductors (order of 100 nm) which lead to higher field at a given drive voltage, and thus lower onset voltage and higher efficiency.

Figure B9.6 shows the level diagram for a (fictitious) double-layer device consisting of an HTL and an ETL. Both layers are assumed to have a bandgap $(I_p - E_a) = 2.5\,\text{eV}$, however, the HTL has lower I_p than the ETL, and the ETL has higher E_a than the HTL.

It is immediately obvious that a single-layer device using either HTL or ETL alone would necessarily have one large (1 eV) injection barrier. In the double-layer architecture both barriers are

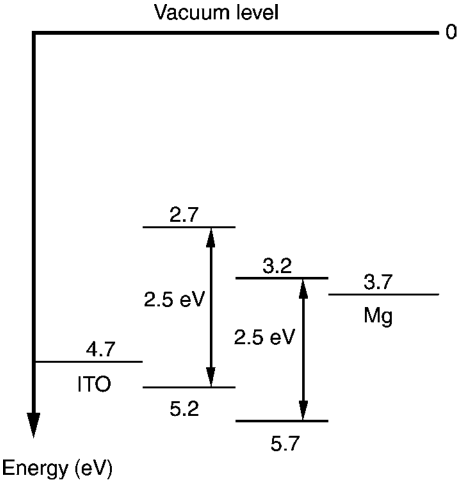


Figure B9.6. Level diagram of a fictitious double-layer device using ITO anode and Mg cathode.

moderate (0.5 eV). Since tunnelling depends exponentially on barrier height, the double layer architecture leads to much improved and more balanced injection.

In addition to the injection barrier, both holes and electrons encounter another internal barrier at the HTL/ETL interface. This additional barrier is not detrimental to device performance though. Instead, it can help to improve the balance between electron and hole currents. Assuming a slightly smaller injection barrier for holes than for electrons, or higher hole mobility than electron mobility, even in a bilayer device we would expect a carrier imbalance with a larger hole current than the electron current. However, since holes encounter an internal barrier, they do not simply cross the device and leave at the cathode as a 'blind' leakage current. Instead, they accumulate at the interface, where they represent a positive space charge.

The effect of the field resulting from that space charge is to improve charge carrier balance. Firstly, it impedes the further injection of majority carriers (holes) from the anode, and secondly, it enhances the injection of minority carriers (electrons) from the cathode. Also, excitons will form at the internal interface, far away from the electrodes. Cathodes in particular have been associated with exciton 'quenching' (i.e. radiationless exciton decay); this is avoided by placing exciton formation at the centre of the device rather than close to the cathode.

As Tang and van Slyke used small molecules, bilayers could readily be manufactured by subsequent evaporation. This approach has been extended to sophisticated multilayer architectures, e.g. by the group of Kido in Yamagata. They have demonstrated some of the brightest and most efficient OLEDs to date (140,000 cd m^{-2} and 7.1% external quantum efficiency) [30].

With polymeric organic semiconductors, vapour deposition is not an option, devices have to be prepared by spincasting instead. Multilayer architectures are harder to realize with spincasting than with vapour deposition, because of the need for 'orthogonal' solubilities. To sidestep solubility problems, in principle, a precursor route may be employed, where the first layer is prepared from a soluble precursor polymer which then is converted in situ into a conjugated and completely insoluble polymer. This has been successfully employed for hole-transporting PPV/electron transporting CN–PPV double-layer polymer OLEDs [31]. However, the precursor route requires lengthy in situ thermal conversion under high vacuum and has generally fallen out of favour with the advent of soluble conjugated polymers.

Recently, a very favourable approach has emerged that combines the ease of injection into a multicomponent device with the simplicity of solution processing. In that approach, a single layer of a blend of a hole-transporting and an electron-transporting conjugated polymer, namely poly(dioctyl fluorene) (PFO) and F8BT, is spincast in one single preparation step. As spincasting implies the very rapid formation of a solid film from solution, the two polymers have little time to phase separate and a solid film may result wherein both polymers remain intimately mixed. Such a mixture has been termed 'bulk heterojunction'. Holes are injected and transported into the (majority component) PFO, but can be transferred easily to F8BT, as it has similar ionization potential. However, F8BT has poor hole mobility due to hole-specific traps. Instead, it has rather high electron affinity and displays comparatively good (albeit dispersive) electron transport [33]. Thus electrons are mobile on the F8BT chain until they encounter a trapped hole. With some further device improvements, highly efficient (4.1 cd A^{-1}) and low onset voltage ($\approx$3 V) OLEDs have been prepared from such blends [34]. The preparation and morphology control of hole/electron transporting blends is the focus of much current research, mainly with a view to photovoltaic applications of organic semiconductors [32].

B9.4.4 Exciton formation

When both hole and electron polarons have been injected into a device, and these drift towards each other under the applied voltage, one expects the formation of hole polaron/electron polaron pairs that are bound to each other. Such bound pairs are termed excitons, and can in some cases be identical to excitons that are formed in an organic semiconductor after the absorption of a photon of light, and subsequent relaxation into the lowest vibrational state of the first excited electronic state. Just as some organic materials display fluorescence (i.e. radiative decay of the excited state), we may find electroluminescence (EL) in such materials.

At first sight, it appears that exciton formation in multilayer architectures is hindered by the internal barrier that carriers of either type encounter at the HTL/ETL interface. However, this is generally not the case. Excitons in organic semiconductors generally display exciton binding energies E_b of a few tenths of an eV [35]. When a carrier has to overcome an internal barrier to form an exciton, it requires a certain amount of energy; however, on exciton formation, E_b is instantly 'refunded'—effectively, the internal barrier is reduced by E_b. Thus, majority carriers remain stuck at an internal barrier and redistribute the internal field in the favourable way discussed earlier, until a minority carrier arrives at the interface. As soon as a minority carrier is available, exciton formation is then helped by the effective barrier reduction E_b. High E_b also stabilizes excitons against dissociation and non-radiative decay.

In 'bulk heterojunction' blends, one carrier has to transfer from one chain to another to form an exciton. This will be the type of carrier for which the energy level offset of either the ionization potentials ($|\Delta I_p|$) or the electron affinities ($|\Delta E_a|$) is smaller. The smaller of the two offsets ($\min(|\Delta I_p|, |\Delta E_a|)$) defines the energetic cost of carrier transfer. Two very different scenarios emerge for the case $\min(|\Delta I_p|, |\Delta E_a|) < E_b$ as opposed to $\min(|\Delta I_p|, |\Delta E_a|) > E_b$. In the former case, formation of excitons from polarons is favoured, while in the latter case, the dissociation of existing excitons into polarons is preferred. In the case of F8/F8BT blends that had been introduced previously, exciton formation is clearly favoured, and such blends are useful for OLED applications. In other hole/electron transport material blends, such as poly(alkyl thiophene)/perylene tetracarboxyl diimide blends, exciton dissociation is favoured, which makes such blends attractive for use in photovoltaic devices [32]. While measurements of $|\Delta I_p|, |\Delta E_a|$, and E_b with sufficient precision to predict exciton formation or dissociation are usually not available, there is a simple experimental approach to decide which is the case: if in a blend fluorescence intensity is much reduced compared to the pure components, excitons are separated efficiently due to the presence of the blend partner.

B9.5 Optimizing efficiency

The discussion in section B9.4 outlines the strategy towards OLED devices with balanced carrier injection and quantitative exciton formation that can be driven at low voltage. The (formidable) challenge that then remains is to maximize the amount of light generated from the excitons. It is obvious that we require a material with a high luminescence quantum yield. The otherwise excellent thiophene-based organic semiconductors fail this criterion and are not suitable for efficient OLED devices. But even given a high luminescence yield, an extraordinary challenge remains that is rooted in the basic properties of excitons.

There is a fundamental difference between the formation of excitons by absorption of a photon, and exciton formation by the combination of electron and hole polarons. The presence of a non-vanishing optical dipole transition moment between a ground state and an excited state that allows for the absorption of a photon implies a difference in the orbital angular momenta of ground and excited states, thus a difference in the orbital quantum number l ('selection rule' $\Delta l = 1$). The overall spin of the resulting exciton, however, will be $S = 0$, just as for the ground state. Such excitons are termed 'singlet' excitons, and correspond to the electron/hole spin combination $(1/\sqrt{2})(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)$. Fluorescence is the re-emission of a photon under return of the singlet excited to the molecular ground state. Unit angular momentum is supplied to the photon from the difference in orbital angular momenta between the singlet excited and the ground state.

When electrically injected electrons and holes combine into excitons, their spins can combine in one of four possible ways. One of those is the singlet combination as discussed as earlier, but there is three more possible polaron spin combinations, namely $(|\uparrow\uparrow\rangle, |\downarrow\downarrow\rangle, (1/\sqrt{2})(|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle))$. These three correspond to so-called 'triplet' excitons with $S = 1$. Triplet excitons have no dipole-allowed (fluorescent) relaxation to the ground state, because there is no orbital angular momentum difference between a triplet excited and the molecular ground state.

Consequently, electroluminescence- and photoluminescence-quantum yields $\eta_{\mathrm{EL/PL}}$ are related via equation (B9.5)

$$\eta_{\mathrm{EL}} = \frac{\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}}}{\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}} + 3}\eta_{\mathrm{PL}} \tag{B9.5}$$

with $\sigma_{\mathrm{S/T}}$ the polaron capture cross-section for singlet and triplet exciton formation, respectively. The assumption that the exciton formation cross-section is independent of the relative orientation of polaron spins ($\sigma_{\mathrm{S}} = \sigma_{\mathrm{T}}$) leads to the prediction that only one-fourth of all formed excitons will be singlet excitons, thus $\eta_{\mathrm{EL}} = (1/4)\eta_{\mathrm{PL}}$. This limit would apply even in the case of an electrically 'ideal' device with ohmic and perfectly balanced injection of electrons and holes, trap free transport, and 'quantitative' exciton formation without leakage currents. Clearly, this limit is undesireable, and the OLED community has devised several approaches to overcome it. Some of these are discussed in the following.

B9.5.1 Enhanced singlet exciton formation

While some experimental studies based on comparisons of EL and PL quantum efficiencies appeared to confirm $\eta_{\mathrm{EL}} = (1/4)\eta_{\mathrm{PL}}$ [36], other studies found high EL quantum efficiencies consistent with a singlet/triplet formation ratio $\approx$1:1 [37], implying $\sigma_{\mathrm{S}} \approx 3\sigma_{\mathrm{T}}$. To determine singlet/triplet formation ratio directly rather than inferring them from EL/PL efficiencies, Vardeny *et al* [38] carried out a systematic magnetic resonance study on a number of organic semiconductors with bandgaps in the visible. They found that $\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}}$ was indeed generally greater than 1, namely between 2 and 5 for different materials. $\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}} \approx (2-5)$ corresponds to $\eta_{\mathrm{EL}} \approx (0.4-0.6)\ \eta_{\mathrm{PL}}$, instead of $0.25\eta_{\mathrm{PL}}$.

Following a later study on an oligomer series [39], Vardeny *et al* now interpret their results in terms of the effective conjugation length of the organic semiconductor, with $\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}}$ increasing with conjugation length. This implies larger $\sigma_{\mathrm{S}}/\sigma_{\mathrm{T}}$ for polymeric than for low-molecular weight organic semiconductors—a finding that potentially can influence the future direction of an entire industry.

Indeed, Friend *et al* found a marked violation of the naïve 1-singlet/3-triplet rule for a polymeric organic semiconductor, but not for a low-molecular weight analogue [40]. They interpret this as the result of the relatively long quantum coherence in a polymeric organic semiconductor. Oppositely charged polarons 'feel' their Coulomb attraction with or without quantum coherence. However, in polymers we have quantum coherence over one conjugation length; consequently, polarons can 'sense' each others' spin from a relatively long distance. Thus, if singlet formation is preferred over triplet formation, polymers have a chance of avoiding triplet formation while electron and hole are still widely separated, i.e. only weakly bound electrostatically. In small molecules, quantum coherence ends at the end of a molecule. Once a hole and an electron are on the same molecule, it is not possible to avoid exciton formation, regardless of spin statistics.

B9.5.2 Electrophosphorescence

As an alternative to the enhanced singlet formation cross-section in polymers, in particular, the low-molecular weight OLED community has developed the concept of 'harvesting' triplets for light emission by using phosphorescence. Phosphorescence bypasses the $\Delta l = 1$ selection rule that normally restricts emissive transitions to singlet excitons. In phosphorescence, the angular momentum necessary for the emission of a photon is supplied from the triplet spin ($S = 1$) rather than from the orbital angular momentum difference of excited/ground state wave functions. To transfer angular momentum from the triplet spin angular momentum to a photon, a spin–orbit coupling term $L{\cdot}S$ is required in the molecular Hamiltonian. Spin–orbit coupling can be of substantial magnitude only if orbitals with higher angular momentum L are present in the molecule. Phosphorescence is therefore typically linked to the presence of atoms with 'high' (i.e. higher than carbon) order number in a molecule ('heavy atom effect'). The phosphorescence transition moment is considerably weaker than for fluorescence, leading to excited state lifetimes typically in the range of microseconds or more ('weakly allowed transition'), as compared to lifetimes in the nanosecond range for fluorescence.

In a typical electrophosphorescent device, a wide bandgap host semiconductor is 'doped' with a small percentage of a phosphorescent emitter. The excitation is transferred from the 'host' to the 'guest' via excitonic energy transfer. Forrest *et al* at Princeton have developed a range of green, yellow, orange and red organoiridium complexes [42], which are exemplified by the particularly efficient red phosphor btp$_2$Ir(acac). When doped into a wide bandgap host, electrophosphorescence with >80% internal quantum efficiency and 60 lm W^{-1} is observed [43]. Using pure btp$_2$Ir(acac) without host matrix has resulted in less efficient devices [44].

Given the fact that polymeric EL devices are considerably more efficient in the formation of singlets than the naïve 1:3 expectation, one may question the need for electrophosphorescence towards enhanced efficiency. However, in particular, in the red, electrophosphorescence is an attractive approach. Due to the response characteristics of the human eye, red dyes must show very narrow emission peaks, otherwise colour purity will be compromised. Iridium-based phosphors display considerably narrower emission bands than typical fluorescent dyes, and are thus particularly useful as red emitters. These phosphors also display relatively short triplet lifetimes (4 μs), thus avoiding problems associated with triplet–triplet annihilation at high brightness, which had been encountered with longer lifetime phosphors [45]. However, electrophosphorescence is ambitious for blue emission due to the need for a high bandgap host semiconductor.

B9.5.3 Organolanthanides

Another approach to 'triplet harvesting' is represented by the organolanthanide dyes. Organolanthanides are somewhat similar to organometallic phosphors, however, the central metal atom is a lanthanide such as europium (Eu) or terbium (Tb). The red dye 'ADS053RE' is a typical example. Organolanthanides owe their properties to the unique electronic structure of the lanthanides (or 'rare earth' metals), which is reflected in their positioning in the periodic table. Up to lanthanium (atomic number 57), the 4f electronic shell remains empty, instead the 5s, 5p, and even the 6s shell are filled first. Only for the elements between cerium (Ce, atomic number 58) and luthetium (Lu, atomic number 71), it becomes energetically more favourable to fill the 4f shell rather than adding more electrons in the sixth shell. The outer (fifth and sixth) shells remain unchanged throughout the rare earths. Hence, all rare earths are chemically very similar. Electronic transitions in the isolated, but incomplete 4f shells are therefore not affected by the chemical bonding.

In a dye such as ADS053RE, the organic ligand can absorb light (typically in the blue or near UV), or can be excited electrically. The exciton is then passed to the central lanthanide and excites an electron of the lanthanide 4f shell; notably this works for both singlet and triplet excitons. The intramolecular excitation transfer is schematically shown in figure B9.7.

Note that the observed emission comes from the radiative decay of the excited 4f state: it is an atomic and not a molecular transition. This is the marked difference between organolanthanides and

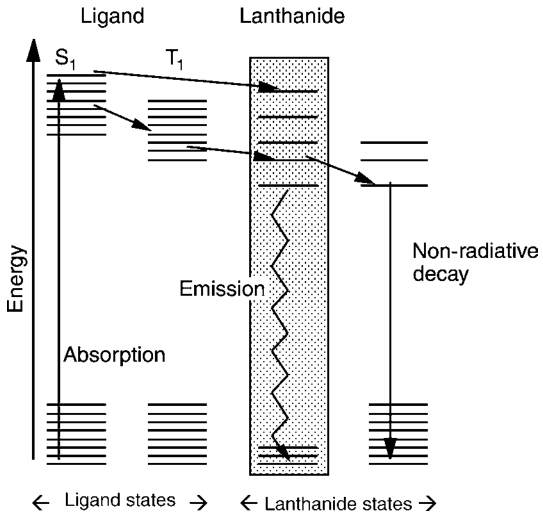


Figure B9.7. An excitation migrates through the molecular (ligand) and atomic (lanthanide) levels in an organolanthanide.

conventional organometallic phosphors. The excited state has a long lifetime of about 1 ms, which makes organolanthanides particularly well suited for passive matrix displays (see chapter C2.4 in this handbook). Nevertheless, due to the localized and isolated nature of the excited state, organolanthanides do not suffer from triplet–triplet annihilation. η_{int} well in excess of 25% can be achieved, and since emission comes from an atomic transition, bands are extremely narrow (FWHM $\approx$ 10 nm). This results in very pure colours (green from terbium (Tb), red from europium (Eu)). Organolanthanide-based OLEDs were developed mainly by Christou *et al* at OPSYS in Oxford [46].

Two major drawbacks for organolanthanide applications are the following. Firstly, for electrical excitation, carriers have to be injected into a rather large bandgap material (deep blue or near UV) even if the emitted light is red. Larger bandgaps generally make good, well-balanced carrier injection harder and lead to reduced power efficiencies and higher onset voltages. Also, as in all transfer-based concepts, the generation of blue light is somewhat problematic. Secondly, the bonding between lanthanide and organic shell has considerable ionic (as opposed to covalent) character. The organic shell acquires a partial negative charge, the lanthanide a partial positive charge. When an electron is injected to the ligand in addition to its partial negative charge, it becomes rather instable against degradation. Consequently, until now, even for encapsulated organolanthanide devices, device lifetimes are poor.

B9.5.4 Conjugated dendrimers

Another recent approach to improved efficiency light emitting devices is the use of dendrons with a conjugated core surrounded by non-conjugated dendrimers. A schematic representation of the dendron concept is given in figure B9.8.

Therein, the conjugated core can be either fluorescent or phosphorescent. The dendron concept seeks to combine the advantages of conjugated polymers and low-molecular weight materials. Dendrimers can be processed from solution and form films in a manner similar to polymers. However, due to the dendronic sidegroups, individual chromophores are shielded from each other. This avoids some of the problems encountered when conjugated polymers are being used. Firstly, conjugated

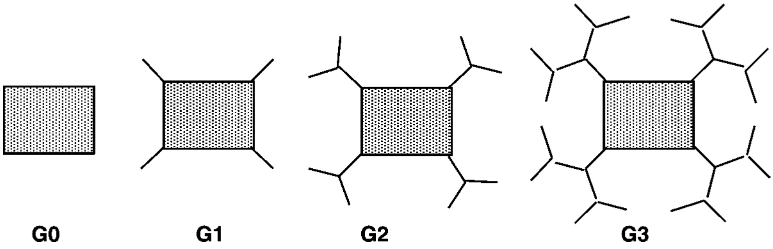


Figure B9.8. Schematic representation of zeroeth- to third-generation conjugated dendrimers. Shaded: conjugated core. Thin lines: dendronic sidegroups (non-conjugated).

polymers often (but not always) display inhomogeneous broadening of emission spectra. This is the result of a statistical distribution of effectively conjugated segment lengths, and is detrimental to colour purity. Secondly, due to interchain interactions like aggregation and excimer formation, quantum efficiency may be reduced, and excimer emission may again compromise colour purity. The major drawback of the dendrimer approach is the much reduced charge carrier mobility due to the increasing separation between conjugated units. Mobility μ scales with the separation between conjugated groups D according to $\mu \sim D^2 \exp(-D/R_0)$.

Samuel *et al* studied dendrimers with a core consisting of three distyrylbenzene groups grouped around a central nitrogen, and dendrimeric sidegroups consisting of *meta*-linked vinylene phenylene groups, up to third generation [47]. OLEDs made from higher generation dendrimers displayed narrow EL spectra that approached solution PL spectra of the conjugated core, and external quantum efficiencies rose steeply with dendrimer generation. This is the result of a successful isolation of the emissive core groups from each other. Carrier injection was not affected by dendrimer generation, however, charge carrier mobility decreased dramatically. For second- and third-generation dendrimers, that did display narrow spectra and improved efficiency, mobility was in the order of only $10^{-8}\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ [48]. Consequently, in transient EL studies under pulsed drive schemes, increasing EL rise times were found for higher generation. Thus, dendrimer-based OLEDs will not be suitable for fast devices suitable for data communication purposes, or for applications that require high current densities (e.g. organic injection lasers). However, the development of dendrimers with phosphorescent cores [49] may lead to OLEDs with very high efficiencies.

B9.5.5 Light outcoupling

After the efficient electrical generation of a photon, this photon must still leave the device to be observed by the viewer of the display. Practically, this outcoupling is often very inefficient, as mentioned in section B9.3.3.

The main loss mechanism here is in-plane waveguiding within the device. To resolve this problem, devices have been designed that suppress waveguiding. With a photoresist-based technique, Samuel *et al* [50] have manufactured a corrugated anode that Bragg-scatters light out of in-plane modes and thus, out of the device and towards the observer. In this way, η_{ext} was improved twofold. A more sophisticated approach for improved outcoupling is to establish a resonant vertical cavity mode in a device. Such cavities have spectrally narrow modes and a strong directional (non-Lambertian) emission characteristic. Thus, spectrally pure colours can be generated even from broadband emitters. The principle of EL from resonant cavities has been established [51], however, device manufacture is difficult. This is due to the need to incorporate a dielectric mirror into the device architecture. Resonant cavities have proven a powerful tool for the investigation of fundamental phenomena in quantum optics of organic semiconductors ('strong coupling' [52]), but for practical devices probably will not be cost efficient. They may, however, play an important role in future developments of organic injection lasers.

B9.5.6 Examples of high performance OLEDs

After discussing several strategies to optimize device performance, here follows a summary description of some of the most efficient OLEDs demonstrated so far and their performance parameters. This list does not claim to be complete and will probably be outdated quickly.

Approach a: conjugated polymer, solution processed.
Device architecture: synthetic metal coated anode, electron blocking layer, emissive blend layer.
Emissive layer: polyfluorene–poly(fluorene-*alt*-benzothiadiazol) blend.
Luminescence data: low onset (3 V and $0.04\,\mathrm{mA\,cm^{-2}}$ for $0.1\,\mathrm{cd\,m^{-2}}$), efficiency $4.1\,\mathrm{cd\,A^{-1}}$ at $16.1\,\mathrm{mA\,cm^{-2}}$ and approximately 1000 cd. Brightness $3500\,\mathrm{cd\,m^{-2}}$ at 12.5 V. Reference [34].

Approach b: small molecule electrophosphorescence, vapour deposited.
Device architecture: multilayer (hole injection layer, emissive layer, electron injection layer).
Emissive layer: Iridium-containing organometallic phosphor doped into electron transporting host layer.
Luminescence data: 19% external quantum efficiency, $60 \pm 5\,\mathrm{lm\,W^{-1}}$ at $1.3\,\mathrm{cd\,m^{-2}}$ and $0.0015\,\mathrm{mA\,cm^{-2}}$. 13.7% external quantum efficiency and $\approx 20\,\mathrm{lm\,W^{-1}}$ at $\approx 1000\,\mathrm{cd\,m^{-2}}$ and $2.1\,\mathrm{mA\,cm^{-2}}$. Reference [43].

Approach c: phosphorescent dendrimer, solution processed.
Device architecture: single-layer blend.
Emissive layer: first generation Iridium-containing dendrimer doped into hole conducting host.
Luminance data: maximum external quantum efficiency 8.1% at 13.4 V, $13.1\,\mathrm{mA\,cm^{-2}}$, giving $28\,\mathrm{cd\,A^{-1}}$ and $3450\,\mathrm{cd\,m^{-2}}$. Maximum power efficiency $6.9\,\mathrm{lm\,W^{-1}}$ at 12 V, $5\,\mathrm{mA\,cm^{-2}}$, and $1475\,\mathrm{cd\,m^{-2}}$. Reference [53].

Approach d: fluorescent low-molecular weight dye, vapour deposited.
Device architecture: multilayer (hole injection layer, hole transport layer, dye-doped emitting layer, electron transport layer).
Emissive layer: aluminium(III)quinolinolato organometallic complex doped with Coumarin 6 laser dye.
Luminance data: $24\,\mathrm{cd\,A^{-1}}$, 7.1% external quantum efficiency, $140{,}000\,\mathrm{cd\,m^{-2}}$ at 12 V. Reference [30].

Approach e: organolanthanide complex, hybrid solution processed/vapour deposited.
Device architecture: double layer, solution-processed blend plus evaporated hole-blocking layer.
Emissive layer: blend of PVK polymeric hole transporter, low-molecular weight PBD electron transporter, and emissive organoeuropium complex.
Luminance data: efficiency $0.73\,\mathrm{cd\,A^{-1}}$ at $23\,\mathrm{cd\,m^{-2}}$ at 16 V drive; brightness $417\,\mathrm{cd\,m^{-2}}$ at $175\,\mathrm{mA\,cm^{-2}}$ and $0.24\,\mathrm{cd\,A^{-1}}$ efficiency at 25 V drive. Reference [46].
Organolanthanide EL with higher efficiency ($2.2\,\mathrm{lm\,W^{-1}}$, $7\,\mathrm{cd\,A^{-1}}$ at $70\,\mathrm{cd\,m^{-2}}$ under 10 V drive) used to be claimed on the webpage of an industrial developer of organolanthanide EL (OPSYS of Oxford, UK), but has now been withdrawn from the webpage.

B9.6 Conclusion

From the performance data listed in section B9.4.6, it is obvious that OLEDs are now a competitive option for display applications, and even room lighting.

However, these performance data need to be put into perspective. A single organic LED is not a display that communicates information. Instead, we require pixellated arrays of OLEDs, which for high-end applications should be able to display large, full-colour pictures at video rates (we are talking about a television screen). Also, one point that has not been discussed here is the all-important issue of device lifetime.

The engineering of efficient, high resolution, full-colour pixellated displays at a competitive price, and approaches to improved lifetimes and their rapid testing via accelerated ageing protocols, are no less formidable challenges than the demonstration of new, highly efficient device concepts. Resolving these challenges requires resources that are generally not available in the academic research environment, and thus were addressed mainly by the industrial players in the field. Contribution C2.4 in this handbook covers device engineering issues from an industrial perspective. From that contribution, it will become apparent that these practical challenges have by and large been addressed and resolved now.

In conclusion, OLEDs will outcompete established display technologies—such as cathode ray tube technology and liquid crystal displays—in the near future.

References

[1] Eley D D 1948 *Nature* **162** 819
Vartanyan A T 1948 *Zh. Fiz. Khim.* **22** 769
[2] Pope M, Kallmann H and Magnante P 1963 *J. Chem. Phys.* **38** 2042
Helfrich W and Schneider W G 1965 *Phys. Rev. Lett.* **14** 229
[3] Partridge R H 1983 *Polymer* **24** 733, 739, 748, 755
[4] Tang C W and van Slyke S A 1987 *Appl. Phys. Lett.* **51** 913
[5] Burroughes J H, Bradley D D C, Brown A R, Marks R N, Mackay K, Friend R H, Burns P L and Holmes A B 1990 *Nature* **347** 539
[6] Hörhold H H, Helbig M, Raabe D, Opfermann J, Scherf U, Stockmann R and Weiß D 1987 *Z. Chem.* **27** 126
[7] Chiang C K, Fincher C R, Park Y W, Heeger A J, Shirakawa H, Louis E J, Gau S C and MacDiarmid A G 1977 *Phys. Rev. Lett.* **39** 1098
[8] Grell M, Bradley D D C, Inbasekaran M and Woo E P 1997 *Adv. Mater.* **9** 798
[9] Sirringhaus H, Wilson R J, Friend R H, Inbasekaran M, Wu W, Woo E P, Grell M and Bradley D D C 2000 *Appl. Phys. Lett.* **77** 406
[10] Redecker M, Bradley D D C, Inbasekaran M and Woo E P 1999 *Appl. Phys. Lett.* **74** 1400
[11] Sheppard J J Jr 1968 Human Color Perception (NY: Elsevier) ISBN 67025430
[12] Baker G L and Pasco S T 1997 *Synth. Met.* **84** 275
[13] Kraft A, Grimsdale A C and Holmes A B 1998 *Angew. Chem. Int. Ed. Engl.* **37** 402
[14] Batchelder D N 1988 *Contemporary Physics* **29** 3
[15] Grimme J, Kreyenschmidt M, Uckert F, Müllen K and Scherf U 1995 *Adv. Mater.* **7** 292
[16] Klärner G and Miller R D 1998 *Macromolecules* **31** 2007
[17] Bernius M, Inbasekaran M, Woo E, Wu W and Wujkowski L 2000 *J. Mater. Sci.: Mater. El.* **11** 111
[18] Groenendaal B L, Jonas F, Freitag D, Pielartzik H and Reynolds J R 2000 *Adv. Mater.* **12** 481
[19] Andreev A, Matt G, Brabec C J, Sitter H, Badt D, Seyringer H and Sariciftci N S 2000 *Adv. Mater.* **12** 629
Yanagi H and Okamoto S 1997 *Appl. Phys. Lett.* **71** 2563
[20] Frey G L, Reynolds K J and Friend R H 2002 *Adv. Mater.* **14** 265
[21] Barashkov N N and Gunder O A 1994 *Fluorescent Polymers* (New York: Ellis Horwood) (ISBN 0133235106)
[22] Bradley D D C 1996 *Curr. Opin. Solid State Mater. Sci.* **1** 789
[23] Bässler H 1993 *Phys. Status Solidi B* **175** 15
Bässler H 2000 *Semiconducting Polymers*, ed Hadziioannou G and van Hutten P F (Weinheim: Wiley-VCH) (ISBN 3-527-29507-0)
[24] Malliaras G G and Scott J C 1999 *J. Appl. Phys.* **85** 7426
[25] Kim J S, Granstrom M, Friend R H, Johansson N, Salaneck W R, Daik R, Feast W J and Cacialli F 1998 *J. Appl. Phys.* **84** 6859
[26] de Blom P W M, de Jong M J M and Vleggar J J M 1996 *Appl. Phys. Lett.* **68** 3308
de Blom P W M, de Jong M J M and Liedenbaum C T H F 1998 *Polym. Adv. Technol.* **9** 390
[27] Horrowitz G 1999 *J. Mater. Chem.* **9** 2021
[28] Warta W and Karl N 1985 *Phys. Rev. B* **32** 1172
Schön J H, Kloc C and Batlogg B, 2001 *Phys. Rev. Lett.* **86** 3843

[29] Campbell A J, Bradley D D C and Lidzey D G 1997 *Appl. Phys. Lett.* **82** 6326

[30] Kido J and Matsumoto T 1998 *Appl. Phys. Lett.* **73** 2866
Kido J and Iizumi Y 1998 *Appl. Phys. Lett.* **73** 2721

[31] Becker H, Burns S E and Friend R H 1997 *Phys. Rev. B* **55** 1

[32] Dittmer J J, Marseglia E A and Friend R H 2000 *Adv. Mater.* **12** 1270

[33] Campbell A J and Bradley D D C 2001 *Appl. Phys. Lett.* **79** 2133

[34] Morgado J, Friend R H and Cacialli F 2002 *Appl. Phys. Lett.* **80** 2436

[35] Bredas J L, Cornil J and Heeger A J 1996 *Adv. Mater.* **8** 447

[36] Baldo M A, O'Brien D F, Thompson M E and Forrest S R 1999 *Phys. Rev. B* **60** 14422

[37] Cao Y, Parker I D, Yu G, Zhang C and Heeger A J 1999 *Nature* 6718 **397** 414

[38] Wohlgenannt M, Tandon K, Mazumdar S, Ramasesha S and Vardeny Z V 2001 *Nature* 6819 **409** 494

[39] Wohlgenannt M, Jiang X M, Vardeny Z V and Janssen R A J 2002 *Phys. Rev. Lett.* **88** art. no. 197401

[40] Wilson J S, Dhoot A S, Seeley A J A B, Khan M S, Köhler A and Friend R H 2001 *Nature* 6858 **413** 828

[41] Sirringhaus H, Brown P J, Friend R H, Nielsen M M, Bechgaard K, Langeveld-Voss B M W, Spiering A J H, Janssen R A J, Meijer E W, Herwig P and de Leeuw D M 1999 *Nature* **401** 685

[42] Lamansky S, Djurovich P, Murphy D, Abdel-Razzaq F, Lee H E, Adachi C, Burrows P E, Forrest S R and Thompson M E 2001 *JACS* **123** 4304

[43] Adachi C, Baldo M A, Thompson M E and Forrest S R 2001 *J. Appl. Phys.* **90** 5048

[44] Adachi C, Baldo M A, Forrest S R, Lamansky S, Thompson M E and Kwong R C 2001 *Appl. Phys. Lett.* **78** 1622

[45] Baldo M A, Thompson M E and Forrest S R 2000 *Nature* **403** 750

[46] Male N A H, Salata O V and Christou V 2002 *Synth. Met.* **126** 7
Moon D G, Salata O V, Etchells M, Dobson P J and Christou V 2001 *Synth. Met.* **123** 355
Christou V, Salata O V and Bailey N J 2000 *Abstr. Pap. Am. Chem. Soc.* **219**, 788-ORGN (Pt 2)
Christou V 2000 *Abstr. Pap. Am. Chem. Soc.* **219**, 99-INOR (Pt 1)

[47] Lupton J M, Samuel I D W, Beavington R, Burn P L and Bässler H 2001 *Adv. Mater.* **13** 258

[48] Lupton J M, Samuel I D W, Beavington R, Frampton M J, Burn P L and Bässler H 2001 *Phys. Rev. B* **63** 155206

[49] Lupton J M, Samuel I D W, Frampton M J, Beavington R and Burn P L 2001 *Adv. Func. Mater.* **11** 287

[50] Lupton J M, Matterson B J, Samuel I D W, Jory M J and Barnes W L 2000 *Appl. Phys. Lett.* **77** 3340

[51] Fisher T A, Lidzey D G, Pate M A, Weaver M S, Whittaker D M, Skolnick M S and Bradley D D C 1995 *Appl. Phys. Lett.* **67** 1355

[52] Lidzey D G, Bradley D D C, Skolnick M S, Virgili T, Walker S and Whittaker D M 1998 *Nature* 6697 **395** 53–55

[53] Markham J P J, Lo S C, Magennis S W, Burn P L and Samuel I D W 2002 *Appl. Phys. Lett.* **80** 2645

B10

Microstructured optical fibres

Tanya M Monro, Anders Bjarklev and Jesper Lægsgaard

B10.1 Introduction

A new class of optical fibre has emerged in recent years: the microstructured fibre [1, 2]. In these fibres the hole-to-hole spacing is typically labelled Λ and d is the hole diameter. Light can be guided using either one or two quite different mechanisms in a microstructured fibre.

The first class of microstructured fibres is the *index-guiding* microstructured fibres. Such fibres are widely known as *holey fibres* (HFs). HFs guide light due to the principle of modified total internal reflection. The holes act to lower the effective refractive index in the cladding region, and so light is confined to the solid core, which has a relatively higher index. Some examples are shown in figure B10.1 (left) and (centre). HFs can be made entirely from a single material, typically pure un-doped silica, although HFs have also been fabricated in chalcogenide glass [3] and in polymers [4]. The effective refractive index of the cladding can vary strongly as a function of the wavelength of light guided by the fibre. For this reason, it is possible to design fibres with spectrally unique properties that are not possible in conventional solid optical fibres. The basic operation of index-guiding fibres does not depend on having a periodic array of holes; in fact, the holes can even be arranged randomly [5].

The optical properties of HFs are determined by the configuration of air holes that forms the cladding region, and HFs can have mode areas ranging over three orders of magnitude by scaling the dimensions of the structure [6]. Small-mode-area fibres can be used for devices based on nonlinear effects [7], whereas the large-mode fibres allow high power delivery [8]. In addition, these fibres can exhibit optical properties not readily attainable in conventional fibres, including endlessly single-mode guidance [9] and anomalous dispersion well below 1.3 μm [10]. Dispersion and birefringence are two properties that can depend strongly on the cladding configuration, particularly when the hole-to-hole separation is small. By exploiting the innate flexibility provided by the choice of hole arrangement, it is thus possible to design fibres with a wide range of characteristics. Note that the modes of all single-material HFs are leaky modes because the core index is the same as that beyond the finite holey cladding, and for some designs this can lead to significant confinement loss [11].

The second guidance mechanism in microstructured fibres can occur if the air holes that define the cladding region are arranged on a strictly periodic lattice. For such structures, photonic bandgaps may appear [1, 12]. These are effective index regions, below the effective cladding index, in which no periodic cladding modes are allowed. By breaking the periodicity of the cladding (e.g. by adding an extra air hole to form a low-index core region), it is possible to introduce a mode that is only allowed in the low-index core region, while being forbidden in the cladding region because of the photonic bandgap. This core mode will, therefore, be guided along the fibre, because of the photonic bandgap of the cladding region. If the core mode has an effective index that is either below or above the effective index range covered by the photonic bandgap at the particular wavelength, the core mode will not be guided.

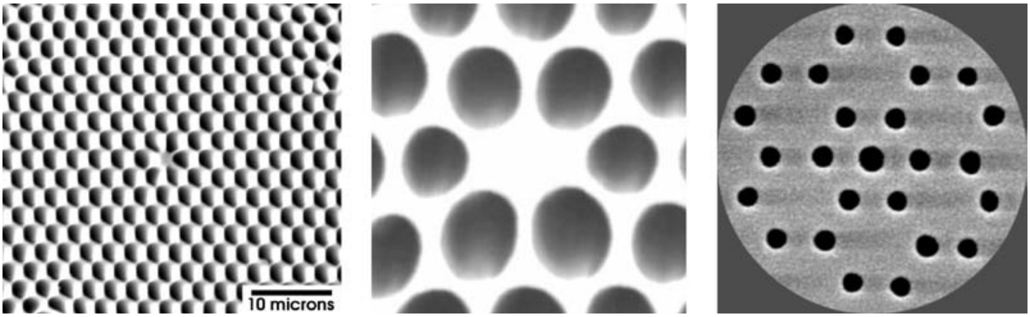


Figure B10.1. Some typical microstructured silica optical fibres. Left: a small-core index-guiding HF (picture supplied by the ORC, Southampton). Centre: a polarization-maintaining index-guiding HF (picture provided by Crystal Fibre A/S). Right: a bandgap-guiding fibre (picture provided by Crystal Fibre A/S).

It has been found that silica–air photonic crystals with air holes arranged in the so-called honeycomb lattice are capable of exhibiting the photonic bandgap (PBG) effect for much smaller air holes than triangular (close-packed hexagonally arranged) photonic crystals [13]. Based on this finding, the first bandgap-guiding microstructured fibres (see figure B10.1) were designed. The inset (right) to figure B10.2 shows how an extra air hole is introduced into the centre of the fibre to act as a low-index defect region.

To understand the waveguiding mechanism, it is valuable to consider the fibre using a modal-index illustration, and figure B10.2 shows such an illustration. It can be seen that two forbidden regions open up by PBG effects below the effective cladding index. No modes appear above the effective cladding index, in agreement with the fact that the low-index core region should not allow guidance of light by modified total internal reflection. However, the extra air hole causes a single mode to be confined to the core defect and correspondingly be guided through the fibre in the frequency range for which the defect

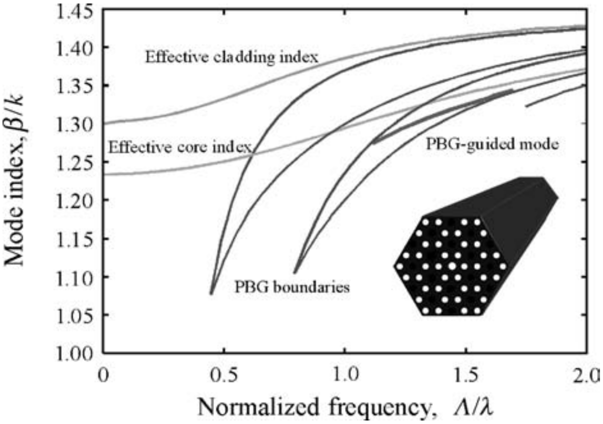


Figure B10.2. Illustration of the two lowest-frequency PBGs of a honeycomb microstructured fibre with a cladding air filling fraction of 30% and a defect hole with same size as the cladding holes. Within the primary PBG, a single degenerate mode is found. This defect mode may not propagate in the cladding structure (due to the PBG effect) and the mode is strongly localized to the region containing the extra air hole that forms the core. The inset shows schematically a honeycomb fibre with the core region formed through the use of an extra air hole.

mode falls within the bandgap. Hence, the PBG fibre may only support guided modes within certain transmission windows. Depending on the extent of the PBGs, these transmission windows may be several microns wide and centred at near-infrared wavelengths [14].

B10.2 Modelling microstructured fibres

The presence of wavelength-scale holes in microstructured optical fibres leads to challenges in the accurate modelling of their optical properties. A wide variety of techniques can be used, ranging from effective step-index fibre models to approaches that incorporate the full complexity of the fibre cross-section. Here these approaches will be reviewed and assessed in terms of their suitability for modelling optical properties of microstructured optical fibres such as their mode area, chromatic dispersion, form birefringence and confinement loss. Some of the issues associated with designing and modelling practical microstructured fibres are discussed.

B10.2.1 Effective index methods

The complex nature of the cladding structure of the microstructured optical fibre does not generally allow for the direct use of analysis methods from traditional fibre theory. However, for index-guiding HFs, a simpler scalar model, based on an effective index of the cladding, has proven to give a good qualitative description of the operation. Birks *et al* first proposed this effective-index approach in 1997 [9]. The fundamental idea behind this work was first to evaluate the properties of the periodically repeated air hole lattice that forms the cladding. By solving the scalar wave equation in a hexagonal cell centred on a single air hole, the propagation constant of the lowest-order mode that can propagate in the infinite cladding material is determined. The hexagonal unit cell shown in figure B10.3 is approximated by a circular one, because a circularly symmetric solution is desired for simplicity reasons. The parameter *r* is the radial distance from the centre of the air hole in figure B10.3. It can be seen that the field is more tightly confined to the silica regions for shorter wavelengths.

This procedure allows the effective cladding index of the fundamental cladding mode to be determined as a function of the wavelength. The next step of the method is then to model the fibre as a standard step-index fibre, employing the strongly wavelength-dependent effective cladding index. The core of the equivalent step-index fibre is assumed to have the refractive index of pure silica, while the core radius is typically taken to be 0.62 times the typical centre-to-centre cladding hole distance Λ. (This assumes that the core is created by the omission of one cladding air hole.)

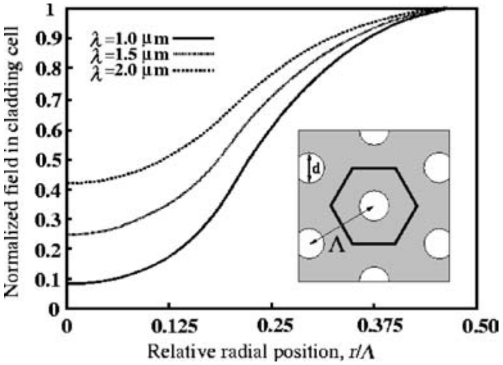


Figure B10.3. Field distribution in a cladding unit cell for three wavelengths when $\Lambda = 2.3\,\mu$m and $d/\Lambda = 0.40$. Figure taken from [15].

Despite ignoring the spatial distribution of the refractive index profiles within HFs, the effective-index method can provide some insight into HF operation. For example, it correctly predicts the endlessly single-mode guidance regime in small-hole HFs. This method has also been used as a basis for the approximate dispersion and bending analysis presented in [16]. However, this reduced model cannot accurately predict modal properties such as dispersion, birefringence or other polarization properties that depend critically on the hole configuration within the cladding. Note that when dispersion predictions are required, the effective-index approach allows for the inclusion of material dispersion properties through the usual Sellmeier formula. It is also noteworthy that recent developments in the fabrication of structures with relatively large air holes have made it relevant to approximate the fibre by an isolated strand of silica surrounded by air [10].

However, one difficulty that arises when using this approach is the question of how to define the properties of the equivalent step-index fibre. One method for making this choice was described earlier. In this work, the core radius was taken to be 0.62Λ: the results obtained using the effective index method were made to agree well with full simulations via appropriate choice of this constant of proportionality. However, for different structures or wavelengths, different choices can become necessary. This restricts the usefulness of this approach, because it is typically necessary to determine the best choice of equivalent structure by referring to results from a more complete numerical model. Reference [17] explores the possibility of choosing the step-index fibre parameters in a more general fashion by allowing a wavelength-dependent core or cladding index. However, to date no entirely satisfactory method for ascribing parameters to the equivalent fibre has been found.

B10.2.2 Structural methods

As mentioned earlier, in order to accurately model HFs, it is typically necessary to account for the complex spatial distribution of air holes that define the cladding. In this section, the techniques that have been developed to do this are outlined. A range of these techniques has also been successfully applied to PBG fibres. Note that when the effective index contrast between core and cladding regions is large, the weak-guidance (scalar) approximation breaks down, leading to inaccurate results, and it is often necessary to adopt a vectorial method that includes polarization effects. This is typically necessary when the air-filling fraction of the cladding is large. Vectorial methods are also appropriate for asymmetric structures. Most of the techniques described in this section can be implemented in both scalar and vector forms.

Much of the early research in microstructured fibres was driven by the desire to fabricate a fibre operating by the PBG effect, which may be obtained in periodically structured material. The idea is that by appropriate choice of crystal structure, the dimensions of the periodic lattice, and the properties of the component materials, propagation of electromagnetic waves in certain frequency bands (the PBG) may be forbidden within the crystal. By packing such periodic material around a fibre core, a PBG fibre with novel propagation characteristics may be fabricated. However, such fibres may not be analysed using simple scalar modelling—or effective-index approaches—because the full vectorial nature of the electromagnetic waves has to be taken into account.

In 1990, the first method for finding PBGs in photonic crystals was described [18]. The method was closely related to methods used for calculating electronic bandgaps in semiconductor crystals, in that it described the magnetic field as a plane wave multiplied by a Bloch function with the two-dimensional periodicity of the photonic crystal. From Maxwell's equations an eigenvalue equation may now be formulated, which is well suited for calculating the PBGs of a periodic dielectric structure, because it describes the field and the structure as a Bloch function [13]. However, to include a core, one has to impose an artificial periodicity, which is handled numerically by creating a supercell with periodically repeated core defects. This yields correct guided solutions, if the supercell is much larger than the guided mode area [19]. Such a supercell approach requires a large number of plane waves, which initiated an interest in models capable of handling many eigenvalues. Note that plane-wave techniques have been used to make a broad range of useful predictions both for PBG-guiding fibres and index-guiding fibres.

Beam propagation methods (BPMs) can also be used to calculate the modal properties of HFs. For example, reference [20] uses a commercial BPM package to investigate a modified conventional germanium-doped fibre in which six large-area holes have been added around the doped core region. A Bragg grating written in the core of the fibre was used to investigate the cladding modes of the structure, and good agreement with the BPM predictions was found. However, because BPM methods calculate the modes of a fibre indirectly by propagating a light distribution along a fibre, they are relatively inefficient computationally.

Another method for analysing microstructured fibres is the application of the finite element method (FEM) as described by Brechet *et al* [21]. In this method, the classical Maxwell differential equations are solved for a large set of properly chosen subspaces, taking into account the continuity of the electromagnetic fields. More specifically, the modelled waveguide is split into distinct homogeneous subspaces of triangular and quadrilateral shapes, the Maxwell equations are discretized for each element, and the resulting set of elementary matrices is combined to create a global matrix system for the entire structure. This method has shown to lead to reliable numerical solutions, and in particular accurate analysis of the modal behaviour in microstructured fibres.

An alternative approach was developed by Mogilevtsev *et al* [22], which describes the modal fields using localized functions. This technique takes advantage of mode localization, and so is more efficient than the plane-wave methods; however, it cannot be accurate unless the refractive index is also represented well.

A hybrid approach, which combines some of the best features of the localized function and plane-wave techniques described earlier, was developed in [6, 23, 24], and some recent extensions to this approach are outlined briefly here. In [6, 23], the air hole lattice is described using a plane-wave decomposition, as in the plane-wave techniques described earlier, and the solid core and the modal fields are described using localized functions. This allows for an efficient description, particularly for idealized periodic structures, because only symmetric terms need to be used in the expansions. In order to model HFs with asymmetric profiles or to obtain accurate predictions for higher-order modes, it is necessary to extend this approach to use a complete basis set, and this was done in [24]. When more complex fibre profiles are considered, the advantages of describing the localized core separately from air holes is diminished, and the best combination of efficiency and accuracy is obtained by describing the entire refractive index distribution using a plane-wave expansion, while using localized functions only for the modal fields. This general implementation of this hybrid approach can be used to explore the full range of HF structures and modes, and can predict the properties of actual HFs by using SEM photographs to define the refractive index profile in the model (see, for example [25]). This allows the deviations in optical properties that are caused by the subtle changes in structure to be explored.

In this implementation, the entire transverse refractive index profile is described using a plane-wave expansion, and the Fourier coefficients are evaluated by performing overlap integrals, which only need to be calculated once for any given structure. The modal electric field is expanded into orthonormal Hermite–Gaussian functions (both even and odd functions are included). These decompositions can be used to convert the vector wave equation into a simple eigenvalue problem (as in the plane-wave method) that can be solved for the modal propagation constants and fields. In order to solve the system, a number of overlap integrals between the various basis functions need to be evaluated. For the choice of decompositions made here, these overlaps can be performed analytically, which is a significant advantage of this approach.

Most of the modelling done to date has considered ideal hexagonal arrangements of air holes. Group theory arguments can be used to show that all symmetric structures with higher than twofold symmetry are not birefringent [26]. However, as the techniques described thus far in this section perform calculations based on a Cartesian grid, they typically predict a small degree of birefringence that can be reduced (but not eliminated) by using a finer grid. However, when modelling asymmetric structures with a form birefringence that is significantly larger than this false birefringence, it is possible to make reasonably accurate predictions for fibre birefringence.

B10.2.3 Predicting confinement loss

As mentioned earlier, for single-material HFs, it can be important to have a means of predicting confinement loss, because all guided modes are intrinsically leaky modes. Even in doped HFs that can have true bound modes, it can be important to understand the leakage characteristics of cladding modes. The confinement loss associated with a given fibre mode can be extracted from the imaginary part of the modal propagation constant. Apart from the BPM approach, none of the techniques described in the previous section can (in their current forms) calculate complex propagation constants.

One technique that has recently been applied to this problem is the multipole approach [11]. This approach is suitable for studying effects caused by the finite cladding region, because it does not make use of periodic boundary conditions. Another advantage of this method is that it calculates the modal fields using decompositions that are based in each of the cladding air holes, and so it avoids the false birefringence problems associated with using a Cartesian coordinate system described earlier. For this reason, this method is also particularly well suited to exploring the symmetry properties of HFs. However, it cannot be used to investigate HFs with arbitrary cladding configurations.

Another technique that has been recently developed that can predict confinement loss is based on representing the refractive index distribution as a series of annular segments [27]. The algorithm uses a polar coordinate Fourier decomposition method with adjustable boundary conditions to model the outward radiating fields. The use of annular segments allows the overlap integrals between the structure and the field components to be performed directly, and so this method can be efficient. In addition, it is possible to represent arbitrary fibre profiles in this way.

In conclusion, a number of techniques have been adopted or developed to model microstructured optical fibres, and a range of novel guidance regimes have been identified in these fibres that promise to lead to a new generation of optical devices with tailor-made optical properties. Many of the techniques described herein complement one another and can be used in conjunction to paint a complete picture of the optical characteristics of any given microstructured fibre. The extremes that are possible in these fibres have highlighted a number of challenges in accurate modelling of their properties, and it seems likely that as the technology for fabricating these structures matures, further challenges will emerge.

B10.3 Highly nonlinear index-guiding HFs

B10.3.1 Nonlinear silica HFs

Design considerations

One of the most promising practical applications of HF technology is the opportunity to develop fibres with a high optical nonlinearity per unit length [7]. The breakthrough for the nonlinear applications of microstructured fibres came with the experimental demonstration of supercontinuum generation in microstructured silica fibres reported by Ranka *et al* [28] in 2000. Among the key elements of this work was that modest optical powers within microstructured fibres can induce significant nonlinear effects. Even though silica is not intrinsically a highly nonlinear material, its nonlinear properties can be utilized in silica optical fibres, if high light intensities are guided within the core. Nonlinear effects can be used for a wide range of optical processing applications in telecommunications and beyond, and examples include optical data regeneration, wavelength conversion, optical demultiplexing and Raman amplification. Consequently, there is great interest in the development of fibres with high values of effective nonlinearity per unit length in order to reduce device lengths and the associated optical power requirements for fibre-based nonlinear devices.

One commonly used measure of fibre nonlinearity is the effective nonlinearity γ [29] which is given by:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\mathrm{eff}}},$$

where n_2 is the nonlinear coefficient of the material ($n_2 \approx 2.2 \times 10^{-20}\,\mathrm{m}^2\,\mathrm{W}^{-1}$ for pure silica), A_{eff} the effective mode area and λ the optical wavelength. For example, standard Corning SMF28 fibre has $\gamma \approx 1\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$. By modifying conventional fibre designs, values of γ as large as $20\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$ have been achieved [30]. This is done by reducing the diameter of the fibre core, and using high germanium concentrations within the core, which both increases the numerical aperture (NA) and enhances the intrinsic nonlinearity (n_2) of the material. Both modifications act to confine light more tightly within the fibre core, and thus increase the nonlinearity γ by reducing the mode area A_{eff}. However, the NA that can be achieved limits the nonlinearity of conventional fibre designs.

Index-guiding HFs can have a significantly larger NA than conventional solid fibre types because the cladding region can be mostly comprised of air. In particular, using HFs with small-scale cladding features and large air-filling fractions (i.e. large values of d/Λ), light can be confined extremely tightly within the core, thus resulting in small mode areas, and large values of γ. Two examples of highly nonlinear silica HFs are shown in figure B10.4. Effective nonlinearities as high as $\gamma \sim 60\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$ have been demonstrated at 1550 nm [31]. This is the best result in terms of nonlinearity reported to date in a pure silica fibre. Note that this result is near to the limit in mode area that can be achieved in pure silica HFs (A_{eff} can be as small as $1.7\,\mu\mathrm{m}^2$ at 1550 nm, see figure B10.6) [32]. Hence nonlinearities more than 50 times higher than in standard telecommunications fibre and two times higher than the large-NA conventional designs are possible. This emerging class of fibre offers an attractive new route towards efficient, compact fibre-based nonlinear devices. Here we review the optical properties of these small-core fibre designs, and present an overview of some of the emerging device applications of this new class of fibres.

Small-core HFs can also exhibit a range of novel dispersive properties of relevance for nonlinear applications [6]. By modifying the fibre profile, it is possible to tailor both the magnitude and the sign of

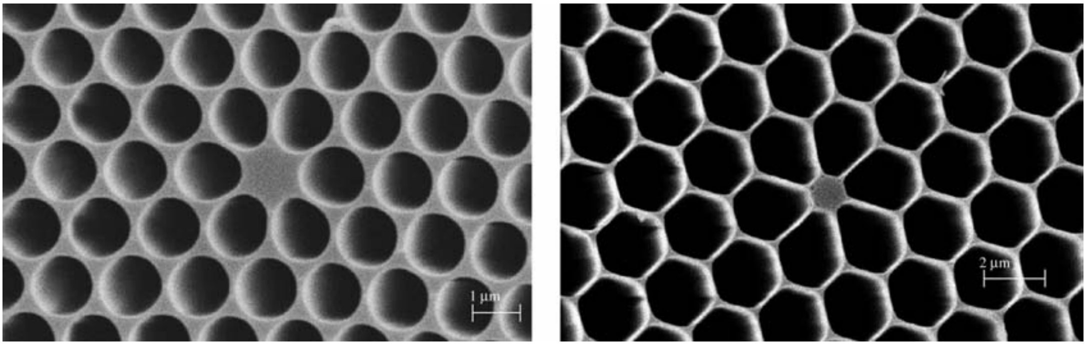


Figure B10.4. Two typical highly nonlinear silica HFs made at the ORC, Southampton. In each case, a small core diameter combines with a large air-filling fraction to result in a fibre that confines light tightly within the solid central core region. Left: $d/\Lambda \approx 0.85$. Right: $d/\Lambda \approx 0.95$.

the dispersion to suit a range of device applications. They can exhibit anomalous dispersion down to 550 nm [10], which has made soliton generation in the near-IR and visible spectrum possible for the first time. An application of this regime was reported in [33], in which the soliton self-frequency shift in an ytterbium-doped HF amplifier was used as the basis for a femtosecond pulse source tunable from 1.06 to 1.33 μm. Shifting the zero-dispersion wavelength to regimes where there are convenient sources also allows the development of efficient super-continuum sources [28], which are attractive for DWDM transmitters, pulse compression and the definition of precise frequency standards. It is also possible to design nonlinear HFs with normal dispersion at 1550 nm [34]. Fibres with low values of normal dispersion are advantageous for optical thresholding devices, because normal dispersion reduces the impact of coherence degradation [35] in a nonlinear fibre device.

Highly nonlinear fibres with zero dispersion at 1.55 μm have long been pursued as these fibres are very attractive for a range of telecom applications such as 2R regeneration [36], multiple clock recovery [37], parametric amplifiers (OPAs) [38], pulse compression [39], wavelength conversion [40], all-optical switching [41] and supercontinuum-based WDM telecom sources [42]. Recently, Hansen *et al* [43] demonstrated a highly nonlinear index-guiding HF with zero-dispersion wavelength at 1.55 μm and a nonlinear coefficient of $18\,\mathrm{W^{-1}\,km^{-1}}$. The fibre was utilized in an all-optical nonlinear optical loop mirror, demultiplexing a bit-stream of $160\,\mathrm{GB\,s^{-1}}$ down to $10\,\mathrm{GB\,s^{-1}}$. This was achieved with only 50 m of fibre, compared to the 2.5 km of dispersion-shifted standard fibre normally required.

Small-core HFs pose a number of challenges for effective modelling. The high index contrast inherent in these fibres necessitates the use of a full-vector method. In addition, any asymmetries or imperfections in the fibre profile, when combined with this large contrast and the small structure scale, can lead to significant form birefringence. Consider, for example, the fibre shown in figure B10.5, which is highly birefringent due to the elliptical shape of the core, with a measured beat length of 0.3 mm, and a polarization extinction ratio of 18 dB at 1550 nm. Using the actual fibre profile in the hybrid orthogonal function method, a beat of length of 0.28 mm is predicted, in good agreement with the measured value. By numerically calculating the deviation of the polarization of the predicted mode from linear polarization, the model predicts an upper bound on the extinction ratio of 19 dB, again in good

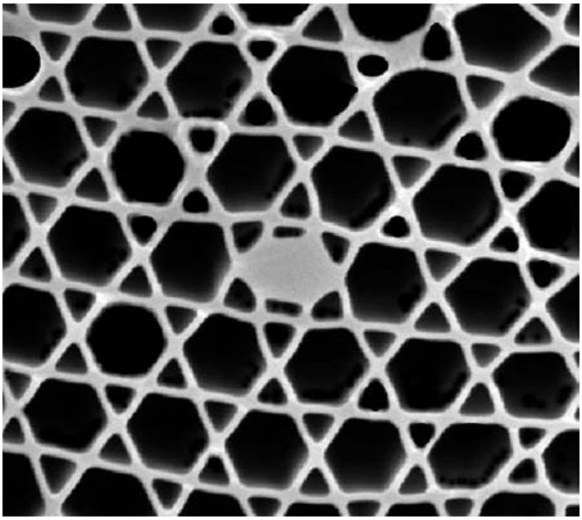


Figure B10.5. Elliptical small core Yb-doped HF (ORC, Southampton).

agreement with observations. In general, even small asymmetries can lead to noticeable birefringence for these small-core fibres. Hence, it is often necessary to use the detailed fibre profile in order to make accurate predictions.

Fibre losses as low as 0.6 dB km^{-1} have now been demonstrated in pure silica HFs [44]. However, when the core diameter is reduced to scales comparable to (or less than) the wavelength of light guided within the fibre, confinement loss arising from the leaky nature of the modes can contribute significantly to the overall fibre loss [32]. Indeed, the small-core HFs fabricated to date are typically more lossy than their larger-core counterparts. In this small-core regime, unless many (6+) rings of holes are used, the mode can *see over* the finite cladding region. Thus, it is important to be able to reliably calculate the confinement loss characteristics of HFs, and the multipole technique [11] has been demonstrated to be a useful tool for exploring the impact of confinement loss in HFs. Here, we briefly outline some general design rules for designing low-loss high-nonlinearity HFs described in [32].

To ascertain the range of effective mode areas that can be achieved using silica glass at 1550 nm, consider the extreme case of a rod of diameter Λ suspended in air. As the diameter of the rod is reduced, the mode becomes more confined, and the effective mode area decreases as shown by the dashed line in figure B10.6. Once the core size becomes significantly smaller than the optical wavelength, the rod becomes too small to confine the light well and the mode broadens again. Hence, there is a minimum effective mode area that, for a given wavelength, depends only on the refractive index of the rod. For silica, this minimum effective mode area is $\approx 1.5\ \mu\mathrm{m}^2$. Figure B10.6 also shows the effective mode area as a function of the hole-to-hole spacing (Λ) for a range of HFs. These HF structures also exhibit a minimum effective mode area due to the same mechanism described. The fibre with the largest air-filling fraction ($d/\Lambda = 0.9$) has a minimum effective mode area of $1.7\ \mu\mathrm{m}^2$, only slightly larger than the air-suspended rod.

As shown in figure B10.6, the hole-to-hole spacing (Λ) can be chosen to minimize the value of the effective mode area, and this is true regardless of the air-filling fraction. However, it is not always desirable to use the structures with the smallest effective mode area, because they typically exhibit higher confinement losses [32]. A relatively modest increase in the structure scale in this small-core regime can lead to dramatic improvements in the confinement of the mode without compromising the achievable effective nonlinearity significantly.

Lower confinement loss and tighter mode confinement can always be achieved by using fibre designs with larger air-filling fractions. Finally, for all fibre designs, it is always possible to reduce

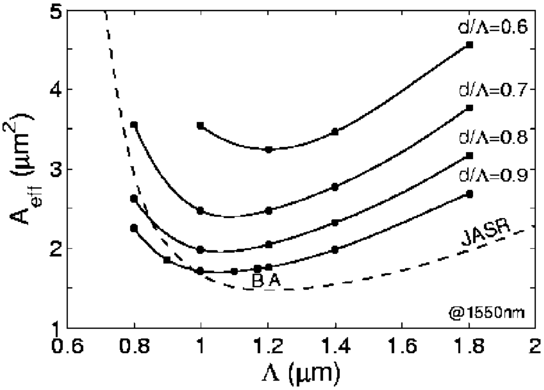


Figure B10.6. Effective mode area (A_{eff}) as a function of the hole-to-hole Λ spacing for a range of fibre designs. JASR corresponds to the case of an air-suspended core of diameter Λ (ORC, Southampton).

the confinement loss by adding more rings of holes to the fibre cladding. In the limit of core dimensions that are much smaller than the wavelength guided by the fibre, many rings (>6) of air holes are required to ensure low-loss operation, which increases the complexity of the fabrication process. With careful design, it is possible to envisage practical HFs with small core areas ($<2\,\mu\mathrm{m}^2$) and low confinement loss ($<0.2\,\mathrm{dB\,km}^{-1}$). Note that although fibre loss limits the effective length of any nonlinear device, for highly nonlinear fibres, short lengths (<10 m) are typically required, and so loss values of the order of $1\,\mathrm{dB\,km}^{-1}$ can be readily tolerated. In addition, note that reducing the core diameter to dimensions comparable to the wavelength of light generally increases the fibre loss for another reason: in relatively small-core fibres, light interacts much more with the air/glass boundaries near the core, and so the effect of surface roughness can become significant [44].

Device demonstrations

A selection of the device demonstrations that have been performed using highly nonlinear silica HFs are reviewed here. The first of these is 2R data regeneration, a function that is a crucial element in any optical network, because it allows a noisy stream of data to be regenerated optically. The first demonstration of regeneration used a silica HF with a mode area A_{eff} of just $2.8\,\mu\mathrm{m}^2$ ($\gamma = 35\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$) at 1550 nm [36]. Typically, devices based on conventional fibres are ~1 km long, whereas in these early experiments just 3.3 m of HF were needed for an operating power of 15 W. Subsequent experiments used an 8.7 m long variant of this switch for data regeneration within an optical code division multiple access (OCDMA) system [45]. Significant improvements in system performance were obtained in this way.

A schematic of the HF-based data regenerator is shown in figure B10.7(*a*). Pulses of light propagating in a highly nonlinear fibre broaden spectrally due to self-phase modulation (SPM), and figure B10.7(*b*) shows the spectrum of 2.5 ps soliton pulses prior to and after propagation through the HF. Figure B10.7(*c*) shows the pulse power that is transmitted through a 1.0 nm narrowband

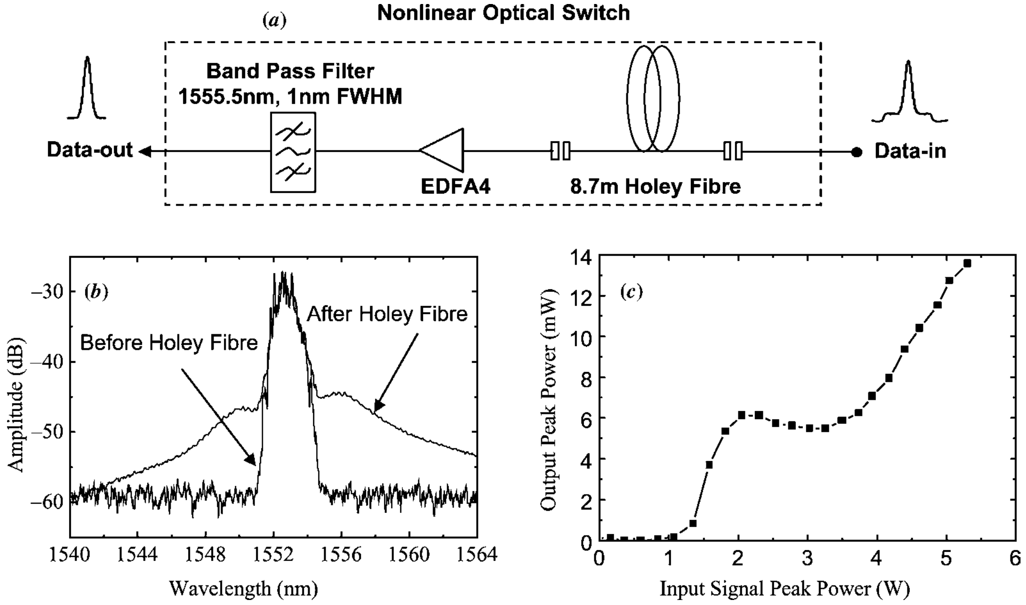


Figure B10.7. (*a*) Schematic of thresholder; (*b*) pulse spectra before/after HF; (*c*) power transmitted (including offset narrowband filter) (work of the ORC, Southampton).

filter (offset by +2.5 nm relative to the incident pulses) as a function of incident pulse peak power. The S-shaped characteristic is suitable for thresholding because at low powers, the pulses do not broaden, and so transmission through the filter is negligible. This corresponds to a '0' in the data stream. For higher powers (~2 W here), substantial SPM occurs, and so transmission through the filter becomes appreciable. This corresponds to a '1'. This device acts to remove noise from an incoming data stream by nullifying all noisy '0' bits, and equalizing all noisy '1' bits.

Fibres with a high effective nonlinearity also offer length/power advantages for devices based on other processes such as Brillouin and Raman effects. The demand for increased optical bandwidth in telecommunications systems has generated enormous interest in the S and L bands, outside the gain band of conventional erbium-doped fibre amplifiers. Fibre amplifiers based on Raman effects offer an attractive route to extending the range of accessible amplification bands. In addition, the fast response time (<10 fs) of the Raman effect can also be used for all-optical ultrafast signal processing applications. Despite these attractions, there is one significant drawback to Raman devices based on conventional fibres: long lengths (~10 km) are generally required, and so Rayleigh scattering ultimately limits their performance. High-nonlinearity fibres offer a method for obtaining sufficient Raman gain in a short fibre length, which eliminates this problem.

For example, reference [46] demonstrated a 70 m fibre-laser-pumped Raman amplifier, and this experiment is outlined in figure B10.8. The amplifier was pumped using a pulsed fibre laser and provided gains of up to 43 dB in the L+ band (figure B10.8(*b*)) for peak powers of ~7 W (figure B10.8(*c*)).

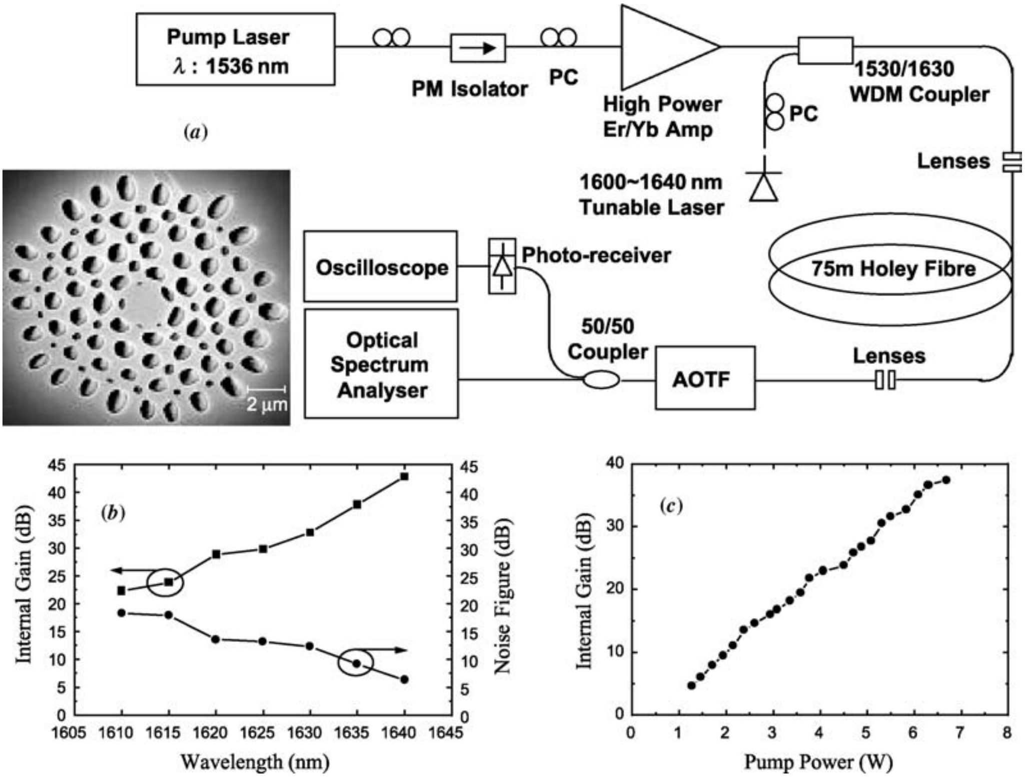

Pump Laser
λ : 1536 nm
PM Isolator
PC
High Power
Er/Yb Amp
PC
1530/1630
WDM Coupler
(a)
Lenses
1600~1640 nm
Tunable Laser
2 μm
Oscilloscope
Photo-receiver
75m Holey Fibre
Optical
Spectrum
Analyser
50/50
Coupler
AOTF
Lenses
(b)
Internal Gain (dB)
Noise Figure (dB)
Wavelength (nm)
0 5 10 15 20 25 30 35 40 45
1605 1610 1615 1620 1625 1630 1635 1640 1645
(c)
Internal Gain (dB)
0 10 20 30 40
Pump Power (W)
0 1 2 3 4 5 6 7 8

Figure B10.8. (*a*) Amplifier schematic (AOTF = acousto-optic tunable filter); (*b*) gain; (*c*) gain efficiency curve at 1635 mm (work of the ORC, Southampton).

Other nonlinear device applications of HFs that have been demonstrated include a CW Raman laser [47] and a WDM wavelength converter [48]. The CW Raman laser was pumped at 1080 nm using a high-power, cladding-pumped Yb-doped fibre laser. The laser had a CW threshold of 5 W, and slope efficiency of 70%. Note the CW power density at the facet ($2\,\mathrm{W}\,\mu\mathrm{m}^{-2}$) demonstrates that HFs can exhibit a good resilience to damage.

Supercontinuum generation in silica-based index-guiding HFs. The high nonlinear coefficient and designable dispersion properties of index-guiding HFs make them attractive for many nonlinear applications of which supercontinuum generation has been the most intensively investigated [28, 49, 50]. The continua have been used in applications like optical coherence tomography [51], spectroscopy and metrology [52]. Supercontinua covering several octaves as well as multi-watt output have been demonstrated [53]. Considerable effort has been made to develop better understanding of the complex interplay of nonlinear processes behind supercontinuum generation, and many of the basic mechanisms (e.g. soliton fission [54, 55], self-phase modulation [56], four-wave mixing and stimulated Raman scattering [49]) are today understood.

An example of the possibilities of supercontinuum generation in index-guiding HFs is presented in figure B10.9 showing an octave-spanning spectrum broadening.

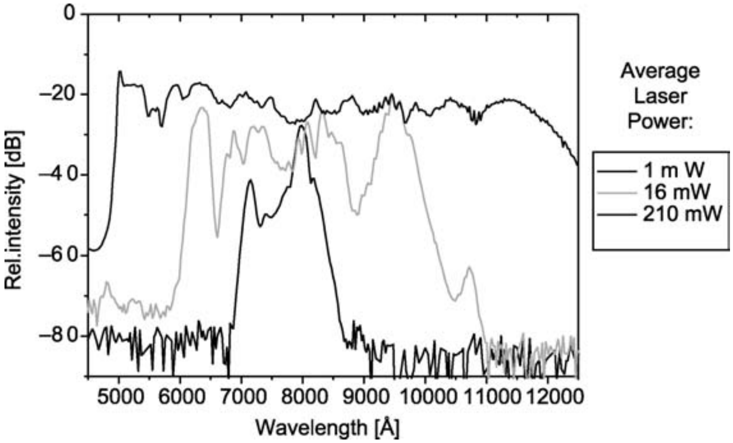


Figure B10.9. Octave-spanning spectrum broadening is made possible with highly nonlinear HFs. This figure illustrates the output from a 50 cm, 2 μm core fibre pumped by a 50 fs 800 nm Ti:sapphire laser. The repetition rate of the pump laser is 76 MHz. The single-mode fibre's zero-dispersion wavelength is 760 nm (work by J.J. Larsen, Aarhus University/NKT Research; fibres provided by Crystal Fibre A/S).

Finally, note that several groups currently are working on detailed theoretical and numerical studies of nonlinear properties of index-guiding HFs. Among the most interesting recent results is the work by Ferrando *et al* [57], which demonstrates self-trapped localized modes in HFs. In this paper, the spatial properties of nonlinearities within the fibres are taken into account.

B10.3.2 Nonsilica HFs

As described earlier, using single-material silica HFs, effective nonlinearities as high as $\gamma \approx 60\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$ are possible. Moving to glasses with a higher refractive index than silica, it is possible to access material nonlinearities that are orders of magnitude larger than that of silica. For example, the chalcogenide glass As_2S_3 has a refractive index of ~2.4 at 1550 nm and is 100 times more nonlinear than silica glass

($n_2(As_2S_3) \sim 2 \times 10^{-18}\,W^{-1}\,km^{-1}$ [58]). The Schott lead-glass SF57 has a refractive index of 1.8 at 1550 nm, and is 20 times more nonlinear than silica ($n_2(SF57) \approx 4 \times 10^{-19}\,W^{-1}\,km^{-1}$ [59]). Recall that for silica, the theoretical lower bound for the effective mode area is ~1.5 μm^2. For the higher index SF57 glass, the minimum effective mode area is reduced to ~0.75 μm^2. Hence, in addition to providing high intrinsic material nonlinearity, such glasses also offer improvements in terms of mode confinement relative to silica.

Conventional fibres made using As_2S_3 have been used to reduce the power levels and fibre lengths required for all-optical switching [58]. Further improvements are to be expected, when such a highly nonlinear glass is combined with the tight mode confinement offered by an HF structure. Although compound glasses are clearly attractive for nonlinear devices, the application of compound glass fibres has been limited because it is difficult to fabricate low-loss single-mode fibres using conventional techniques. Single-material fibre designs avoid core/cladding interface problems, and so should potentially allow low-loss fibres to be drawn from a wide range of novel glasses. As we demonstrate here, HF technology provides a powerful new technique for producing single-mode compound glass fibres. We briefly review some recent results using SF57 glass from [60].

Like many compound glasses, SF57 has a low softening temperature (~520 °C) and so it is possible to extrude the HF preform directly from bulk glass, and the cross-section of a fibre fabricated from an extruded preform is shown in figure B10.10.

The core diameter is ≈2 μm and the core is suspended by three ≈2 μm long supports that are less than 400 nm thick. These supporting struts allow the solid core to guide light by helping to isolate the core from the outer solid regions of the fibre cross-section. Although single-material fibres support only leaky modes, it is possible to design low-loss fibres of this type (see [27]). This can be done by ensuring that the supporting struts are long and fine enough that they act purely as structural members that isolate the core from the external environment.

Extrusion offers a controlled and reproducible method for fabricating complex structured preforms with good surface quality. In addition, extrusion can be used to produce structures that could not be created with capillary stacking approaches, and so a significantly broader range of properties should be accessible in extruded HFs.

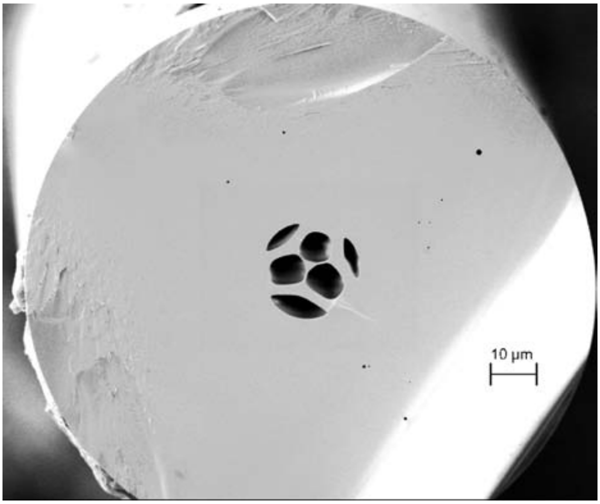


Figure B10.10. (*a*) SEM of the extruded SF57 microstructured fibre from [60]. Picture supplied by the ORC, Southampton.

The measured effective nonlinearity for this fibre is $\gamma \approx 550\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$, and assuming the value of n_2 for this material given earlier, this implies that $A_{\mathrm{eff}} \approx 3.0 \pm 0.3\,\mu\mathrm{m}^2$. This mode area is ≈30 times smaller than conventional SMF28 fibre, and the intrinsic material nonlinearity of the SF57 glass provides an n_2 that is ≈20 times larger than silica, and so the effective fibre nonlinearity is more than 500 times larger than that of standard single-mode optical fibre. These results represent the first demonstration of single-mode guidance and nonlinearity enhancement in a nonsilica HF.

Possibilities exist to increase the nonlinearity still further through the use of glasses with still higher nonlinearity than SF57 (such as As_2S_3), and to design fibres with smaller cores. Such fibres promise a route towards record effective fibre nonlinearities, paving the way to nonlinear fibre devices with unprecedentedly low operating powers (1–10 mW) and remarkably short device lengths (0.1–1 m).

B10.4 Large-mode-area index-guiding HFs

B10.4.1 Passive fibres for high power delivery

Index-guiding HFs offer an alternative route towards large mode areas [8]. The development of large-mode holey fibres (LMHFs) is important for a wide range of practical applications most notably those requiring the delivery of high-power optical beams. For many of these applications, spatial mode quality is a critical issue and such fibres should preferably support just a single transverse mode. Relatively large-moded single-mode fibres can be made using conventional fibre doping techniques such as modified chemical vapour deposition simply by reducing the NA of the fibre and increasing the fibre core size. However, the minimum NA that can be reliably achieved is restricted by the accuracy of the control of the refractive index difference between the core and cladding. In addition to offering large mode areas, LMHFs offer other unique and valuable properties; most notably they can be single moded at all wavelengths. This is in contrast to standard optical fibres, which exhibit a cut-off wavelength below which the standard fibre becomes multimoded. HF technology also provides an alternative, potentially more accurate route to controlling the index difference between core and cladding regions of the fibre.

LMHFs can be produced by designing fibres with a large hole-to-hole spacing ($\Lambda > 5\,\mu\mathrm{m}$) and/or small air holes ($d/\Lambda < 0.3$). Figure B10.11 shows the near-field mode profile in a typical LMHF superimposed on a scanning electron microscope (SEM) picture of the fibre profile. This fibre is interesting not only because it has a very large mode field diameter (MFD around 20 μm at a wavelength of 1550 nm), but also because it is single moded for any wavelength at which the silica host material is transparent.

Because LMHFs rely on a very small effective index contrast between core and cladding, the fibres can be sensitive to macro-bending, and this is discussed further later in this section. For this reason, the fibre properties can be highly sensitive to the precise details of the fibre structure, and so accurate fabrication techniques are required. Figure B10.12 shows cross-sectional pictures of two typical LMHFs with regularly arranged air holes.

The models described in section B10.2 can be applied to model the optical properties of these fibres, although typically extra care is needed due to the wide range of spatial scales present. Polarization effects are typically less important in this class of fibres, and it is often sufficient to use a scalar model.

Macroscopic bend loss ultimately limits the practicality of such large-mode fibres, and so understanding bend loss is important in the design of this class of fibre. Two distinct bend loss mechanisms have been identified in conventional fibres: transition loss and pure bend loss [61]. As light travels into a curved fibre, the mode distorts, causing a transition loss (analogous to a splice loss). Pure bend loss occurs continually along any curved section of fibre: at some radial distance (r_c in figure B10.14(*d*)), the tails of the mode need to travel faster than the speed of light to negotiate the bend, and are thus lost.

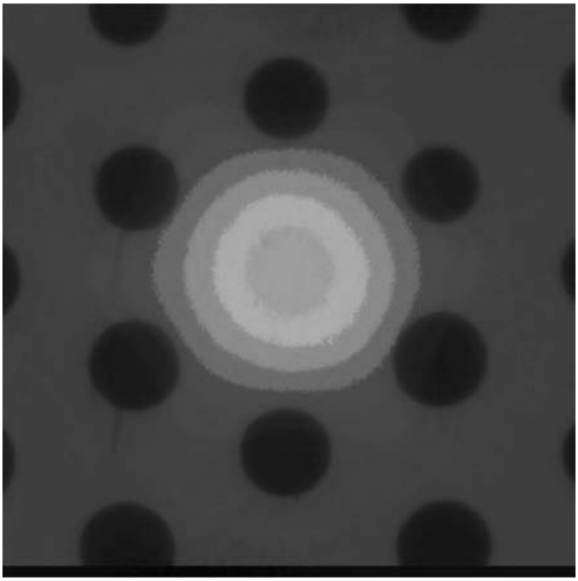


Figure B10.11. Cross-section of a large-mode-area HF with a mode field diameter of 20 μm at 1550 nm. The near-field picture of the guided mode is superimposed. Photograph provided by Crystal Fibre A/S.

Using the simplified effective index method described in section B10.2, it is possible to derive useful formulas for the pure bend loss in these fibres by applying standard results for the power loss coefficient of standard step-index fibres (see [62, 63]). Utilizing this approach, the pure macro-bending loss can be derived from the coefficients of the Bessel function of the equivalent step-index fibre, as done in [64]. Figure B10.13 illustrates the bending radius dependence of the operational windows for a specific LMHF with air hole diameter $d = 2.4\,\mu\text{m}$ and hole-to-hole spacing $\Lambda = 7.8\,\mu\text{m}$. As can be seen in

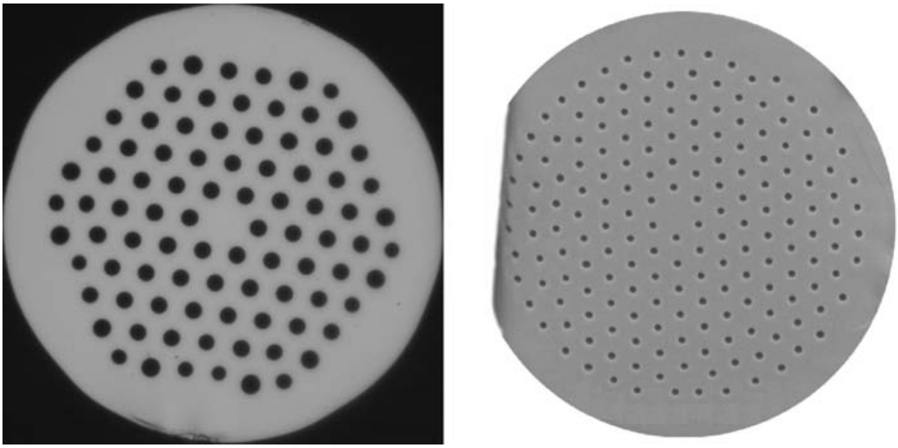


Figure B10.12. Full cross-section of two typical large-mode-area HFs with different air-filling fractions (both have core diameters of 15 μm). Left: photograph provided by Crystal Fibre A/S. Right: photograph provided by the ORC, Southampton.

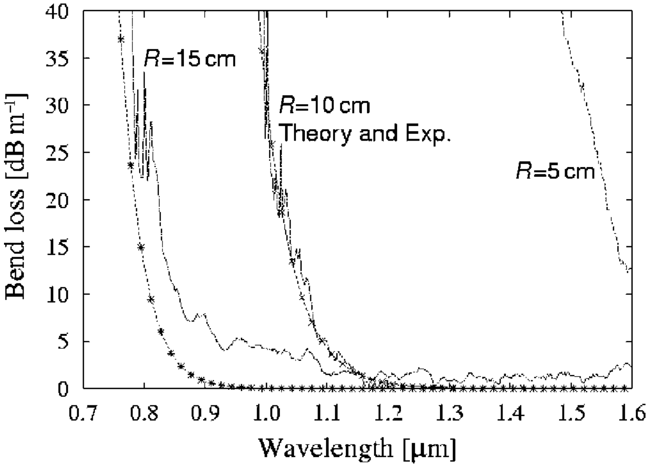


Figure B10.13. Calculated and measured values of spectral bending loss of LMHF. For operation around 1.5 μm, the critical bending radius is ~6 cm (from [64]).

Figure B10.13, a short-wavelength loss edge is evident for LMHFs. This is in contrast to the standard fibre case, where only a long-wavelength loss edge is found. For so-called endlessly single-mode LMHFs—single-mode fibres that can be designed with a very large mode area—it is, therefore, important to notice that macro-bending losses in practice will limit the operational wavelength.

Figure B10.13 further includes the predictions of the effective-index model, and as can be seen, the model is capable of predicting accurately the spectral location of the short-wavelength bend loss edge [8]. The long-wavelength bend loss edge of this specific fibre is positioned at mid-infrared wavelengths for all bending radii. Although this figure presents results for one particular LMHF, a number of generalizations can be made. Unsurprisingly, the critical bending radius and the spectral width and location of the operational window depend strongly on hole size (d) and hole-to-hole spacing (Λ). Generally, larger holes result in broader operational windows, whereas the hole-to-hole distance roughly determines the centre position of the window (as a first approximation the minimum bend loss occurs at a wavelength around $\Lambda/2$) [64]. Hence, standard telecommunications wavelengths fall on the short-wavelength loss edge for large-mode HFs. Despite this, LMHFs have been demonstrated to possess comparable bending losses to similarly sized conventional fibres at 1550 nm [65].

The results described earlier have focused on pure bend losses, which will be the dominant form of bend loss in long fibre lengths. For shorter lengths of fibre, it can be important to investigate the impact of transition loss too. Transition and pure bend losses can be distinguished experimentally by progressively wrapping a fibre around a drum of radius R_0 [61]. The fibre experiences a sharp change of curvature as it enters and leaves the drum surface, which results in transition losses at these points. As the angle is increased, the length of the curved section (and the pure bend loss) increases linearly. Figure B10.14 shows the measured loss as a function of angle for $R_0 = 14.5$ mm (taken from [66]). Each data set shows two regions: the curved section of each plot is the transition region, while the pure bend loss dominates as the length of the bent fibre increases. Results for two different angular orientations of the same fibre illustrate that the geometry of the cladding structure has a noticeable effect on the bend loss characteristics. Hence, in order to understand and predict such effects, it is necessary to use a numerical method that accounts for the complex structure, because effective-index methods cannot account for orientationally dependent behaviour.

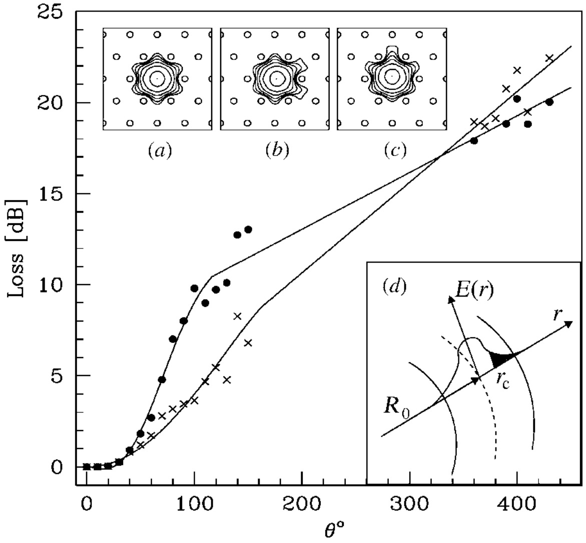


Figure B10.14. Loss as a function of angle for two fibre orientations (see text and reference [66]) for the HF in figure B10.12 (right). (*a*) Calculated mode profile for this HF; (*b*) and (*c*) calculated mode for a bend in horizontal and vertical directions, respectively (contours every 2 dB); (*d*) slice through mode propagating around bend of radius R_0.

One way of modelling the propagation of light in a fibre with a radius of curvature R_0 is to scale the refractive index using the transformation: $(1 + (2r\cos\alpha)/R_0)^{1/2}$, where the coordinates are defined in [67]. Using this transformation, the modes of the bent microstructured fibre can then be calculated using one of the full-structural techniques described in section B10.2 [66]. Note that the slant introduced by this transformation cannot be described in all of the models. To calculate the modes shown in figure B10.14 the hybrid orthogonal function model was used: modal intensities are shown for (*a*) a straight fibre, and (*b*) and (*c*) for fibre bent in horizontal and vertical directions, respectively. Using these modes, the transition loss can be calculated as the overlap between the straight and bent modes, and for pure bend loss it is assumed that the fraction of energy in the guided mode at $r > r_c$ is lost over some distance scale. This model gives good quantitative agreement with experimental data, and allows trends relating to the angular orientation of the fibre to be identified.

B10.4.2 Active large-mode-area fibres

The capillary-stacking techniques that are generally used to make LMHFs can be readily adapted to allow the incorporation of high-NA air-clad inner claddings within jacketed all-glass structures [68]. This technique was described by DiGiovanni *et al* [69], who outlined how an outer cladding having a very high air-filling fraction may result in an effective refractive index below 1.35. The optical characteristics of the light guided within the core of the optical fibre are essentially independent of the second outer cladding, and the fibre can become insensitive to the external environment. The advantage in this connection is the large index step between air and silica, which makes microstructured fibres with extremely high NA values (compared to standard fibres) possible. Fibres with NAs as high as 0.9 have been demonstrated. High-NA fibres (typically multimode) collect light very efficiently from a very broad space angle and distribute light in a broad angle at the output end. They could find use for pig-tailing broad-area-emitting lasers and for lighting applications such as windmill warning signals and endoscopy. An example of a high-NA multimode fibre is shown in figure B10.15.

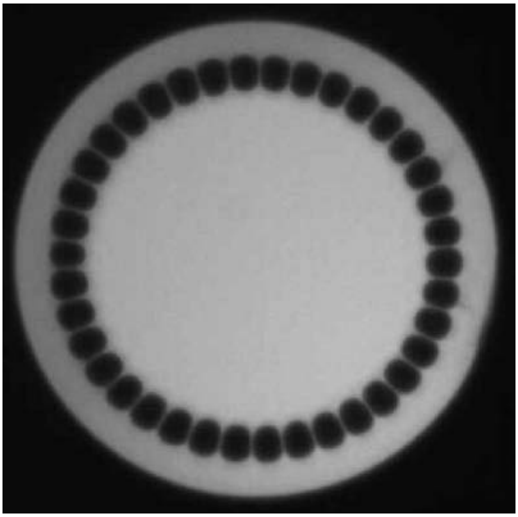


Figure B10.15. Cross-section of a pure silica fibre with a 25 μm multimode core diameter. The NA is higher than 0.55. Photograph provided by Crystal Fibre A/S.

The potential for high-NA fibres is also of great interest in the context of high-power, rare-earth-doped (e.g. Yb^{3+}, Nd^{3+}), LMA devices. The use of such rare-earth dopants in LMHFs is, however, challenging. The presence of dopants (and associated co-dopants such as germanium, aluminium and boron that are required to incorporate the rare-earth ions at reasonable concentrations and to maintain laser efficiency) modifies the refractive index of the host glass. This affects the NA of the fibre, and can lead to the loss of some of the most attractive LMA HF features, such as broadband single-mode guidance, unless care is taken in the fibre designs.

HF preforms are typically created by stacking capillaries around a solid rod, which ultimately forms the core. To produce active HFs, it is thus necessary to produce a doped core-rod, which can be achieved by extracting the core region from a conventional doped fibre. In the example given in [70], the starting point was an aluminosilicate ytterbium-doped rod with an NA of 0.05 produced using conventional MCVD techniques. In addition, it is straightforward to adopt HF fabrication technology to achieve an all-glass double-clad structure, which is advantageous for the efficient use of low-brightness pump sources. In [70], a low-index outer-cladding region is formed simply by inserting thin-walled capillaries into the preform stack around the relatively thicker-walled capillaries used to define the inner holey cladding region. An SEM photograph of a fibre produced in such a manner is shown in figure B10.16. The inner-cladding NA was measured to be 0.3–0.4 in a short piece (~10 cm) of the fibre. In this design, the index difference introduced by the dopants was small relative to the core/inner cladding index contrast, and so the structure remained single mode at the operation wavelength. An offset core was used to break the symmetry of the cladding, and so to enhance the pump absorption.

Both core- and cladding-pumped lasers have been realized based on the air-clad ytterbium-doped HF shown in figure B10.16, and these results are reviewed briefly here.

Core-pumped LMHF laser

A single-transverse-mode Ti:sapphire laser operating at 976 nm was used as a pump source to examine the laser performance of the fibre. A conventional Fabry–Perot cavity configuration was formed by the 4%

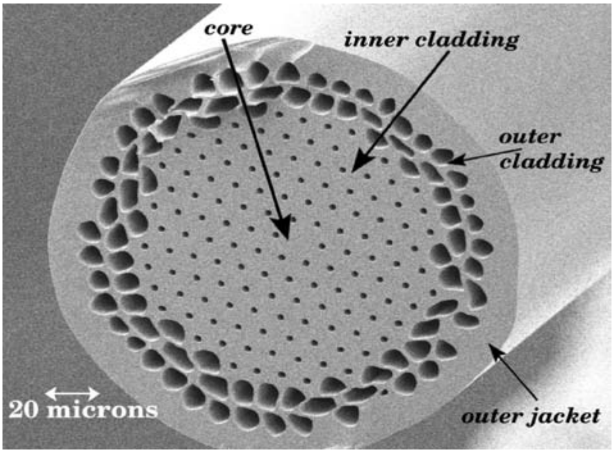


Figure B10.16. SEM of the air-clad Yb-doped HF from [70]. Picture supplied by the ORC, Southampton.

Fresnel reflection from the launch end of the fibre and a lens-coupled high reflector placed at the other end of the fibre. Slope efficiencies as high as 82% were recorded, comparable to the best conventional ytterbium fibre lasers. As expected, the output beam was observed to be robustly single mode.

Cladding-pumped LMHF laser

The pump was a low-brightness fibre-coupled laser diode at 915 nm. Using again a simple Fabry–Perot cavity with 4% feedback, average powers in excess of 1 W were achieved in a 7.5 m long fibre with a slope efficiency of 70%. As well as optically isolating the inner structure from the external environment, the air cladding also thermally isolates the laser. Although it might be imagined that this could lead to thermal problems under high-power operation, no such problems were encountered even at multi-watt pump levels. Q-switching and mode-locking were also reported using this cladding-pumped laser. In the Q-switching experiment, we obtained ∼50 J stable pulse at repetition rates of a few kilohertz. The output pulse duration was ∼10 ns, and the corresponding peak power was 5 kW. In the mode-locking experiments, fundamental mode-locking was obtained over a wavelength tuning range of 60 nm. The pulse duration was estimated to be ∼100 ps. An output power of more than 500 mW was achieved for a pump power of 1.33 W. Ultimately, it should be possible to develop compact ∼multi 10 nJ femtosecond pulse sources operating at 1 μm using LMHFs in conjunction with the well-established Kerr nonlinearity based stretched pulse mode-locking technique.

The primary advantages of these forms of fibre relative to conventional polymer-coated dual-clad fibres are that they will ultimately allow for all-glass structures, with larger inner-cladding NAs (at least >0.5) and good pump mode mixing. In addition, they offer single-mode guidance in cores that are at least as large (but most likely larger) as those than can be made in conventional fibres. In device terms, these features will translate to advantages including, amongst others, the possibility of higher coupled diode powers (for a given cladding dimension), shorter device lengths and extended tuning ranges.

B10.5 PBG fibres

While the index-guiding HFs discussed in the previous sections can to a first approximation be regarded as a variant of the traditional step-index fibre with a larger and wavelength-dependent index contrast between core and cladding, the PBG fibres constitute a fundamentally different class of waveguides. In PBG fibres, light is not to a region of higher refractive index, but is rather localized at a defect placed in a PBG material, which suppresses transverse propagation at the frequency of the guided mode. Due to this fundamental difference, the PBG fibres attracted considerable academic interest at an early stage and several kinds of PBG fibre were realized experimentally within 2–3 years after the first successful fabrication of an index-guiding microstructured optical fibre [12, 71]. Subsequently, however, most of the research and development effort within the field has focused on the index-guiding fibres due to their great and immediate potential for practical applications. As this technology matures, it can be expected that a greater proportion of academic and industrial research effort will again be directed towards the case of PBG fibres in order to elucidate their potential applications. In this section, we summarize current knowledge about the two PBG fibre types that have been realized experimentally: *honeycomb fibres*, whose properties in many respects resemble those of index-guiding microstructured fibres, and *air-guiding fibres*, which can only be realized through the PBG effect.

For the PBG-guiding mechanism to operate, the bandgap must extend over the whole plane perpendicular to the direction of propagation. In silica–air structures, such a gap can only be achieved for a finite longitudinal propagation constant, the minimum value required being dependent on the fibre structure. The simple triangular arrangement of cladding holes (commonly used for making index-guiding microstructured fibres) only provides a complete bandgap for high air-filling fractions of 30% or higher. In contrast, arranging the holes on a honeycomb lattice makes it possible to open up complete gaps at much lower fill fractions (<1%) [72]. For this reason, the first PBG fibre experimentally fabricated was based on the honeycomb lattice. The basic design and dispersion properties of these fibres are illustrated in figures B10.17 and B10.18: the perfect honeycomb lattice can be thought of as an array of silica rods, separated by rings of air holes. The core region of the waveguide is created by introducing an extra air hole in the central silica rod, thereby lowering the effective index of this region. It is a general feature of PBG-guiding fibres that the central defect is created by lowering the effective index of the core region, whereas for index-guiding microstructured fibres, the effective core index is raised to trap the guided mode. The magnitude and position of the bandgaps are controlled by the radius of the cladding holes. Depending on the size of the central core hole, the fundamental defect mode may be pushed into the first bandgap, into the higher-order bands as a leaky resonance state, or, as is the case for the structure shown, into a higher-order (here the second) bandgap. In principle, any of the bandgaps may

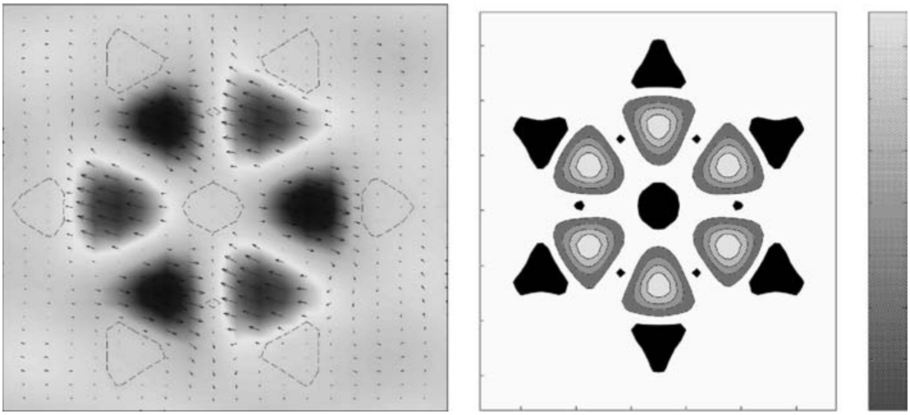


Figure B10.17. Poynting vector (left), and magnetic field (right) and distribution of a guided third-order mode with a wavelength of 500 nm in a honeycomb PBG fibre (black regions on right picture are air holes). The pitch is 2 μm and the air-filling fraction of the cladding is 5.3%. Illustrations from [75].

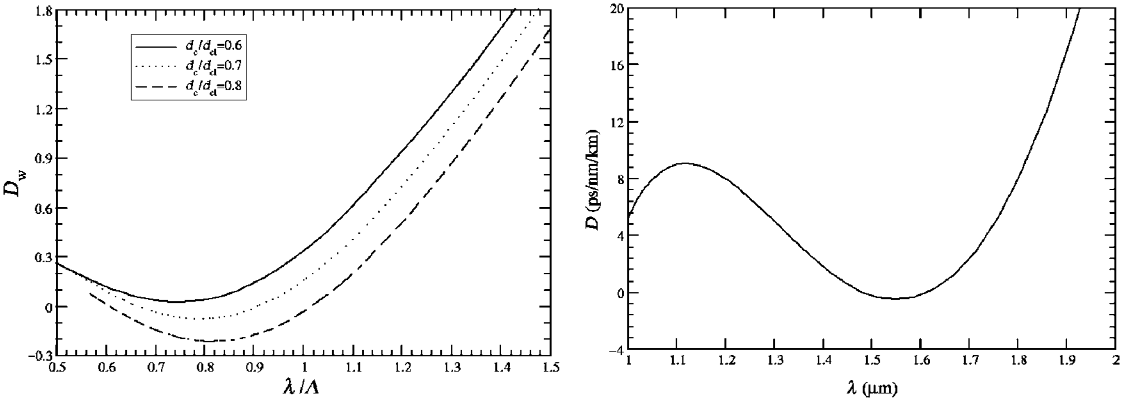


Figure B10.18. Left: normalized waveguide GVD for a honeycomb PBG fibre with different core hole sizes. Right: GVD of a honeycomb fibre designed to have small dispersion near 1.55 μm.

support a number of guided modes. By proper tuning of the core and cladding hole sizes one can, therefore, create fibres which are either single- or multimode, and in contrast to the case of index-guiding fibres, the single-mode fibres may have higher-order bound modes, while the fundamental mode is not guided. In fact, the honeycomb PBG fibre fabricated by Knight *et al* [71] had a guided third-order mode which was transmitted over a 5 cm fibre length, whereas no fundamental mode was found to be guided. The measured properties were in good agreement with calculations performed by Broeng and co-workers. Examples of the confined field patterns calculated are shown in figure B10.17.

While the fabrication of this fibre constitutes an important proof-of-principle of the PBG waveguiding mechanism for honeycomb fibre designs, it is to be expected that practical applications of PBG fibres as functional devices are likely to be based on fibres supporting a fundamental guided mode. Although such fibres have not yet been fabricated, however, their basic properties are known from theoretical studies.

The waveguide group velocity dispersion (GVD) of the honeycomb fibres with a central core hole is mostly anomalous [73], but regions of normal waveguide GVD do occur for cladding hole sizes greater than $\sim 0.5\Lambda$ (where Λ in this connection is defined as the distance between adjacent cladding holes). An example of waveguide dispersion behaviour for a structure with a cladding hole diameter, d_{cl}, of 0.56Λ and various core hole diameters, d_{c}, is shown in figure B10.18 (left). The dispersion is reported in units of $1/(ca)$ where a is the lattice period of the periodic cladding and c the velocity of light in vacuum. It may be noted that the possibility of varying the size of the core hole provides a degree of design freedom.

The shapes of the waveguide dispersion curves imply that the honeycomb PBG fibres have roughly the same potential as the index-guiding fibres with respect to dispersion engineering: chromatic GVD curves may be flattened over large wavelength intervals, zero dispersion may be obtained at small wavelengths, and fibres with large anomalous GVD coefficient (several hundred ps nm^{-1} km^{-1}) may be fabricated. Figure B10.18 (right) shows the chromatic GVD curve for a fibre with $d_{\mathrm{cl}}/\Lambda = 0.64$ and $d_{\mathrm{c}}/d_{\mathrm{cl}} = 0.55$ (d_{c}, d_{cl} being the diameter of core and cladding air holes, respectively, and Λ the pitch). The fibre has been designed such that both the dispersion coefficient and its first derivative go to zero around $\lambda = 1.55$ μm, to demonstrate the great design flexibility of this fibre type.

While the ordinary honeycomb fibre design depicted in figure B10.2 may be shown to be free of birefringence, one can imagine many ways of breaking the sixfold symmetry of the structure, and the possibility of fabricating polarization-maintaining PBG fibres has been investigated theoretically [74].

In this work, two out of six air holes in the innermost ring surrounding the core defect were varied in size. A variety of hole sizes were considered, and a birefringence as high as 2×10^{-3} (the difference in effective index of the two polarization modes) was obtained. A different approach, which keeps the cladding structure symmetric, but introduces an elliptical core defect, was investigated in [75]. By using a cladding structure with large air holes ($d/\Lambda = 0.7$), it was found that one polarization state could be pushed out of the PBG thus creating a truly single-moded fibre. This possibility is another unique features of fibres guiding by the PBG effect.

So far, most of the characteristics of honeycomb PBG fibres found by theoretical investigations appear similar to those of index-guiding microstructured fibres. However, there is one possibility which is unique to fibres guiding by the PBG mechanism, namely that of creating structures which predominantly guide the light in air. This is possible if a cladding structure can be fabricated in which modes with an effective index of 1 fall within the bandgap, or, in other words, where the effective index of the cladding mode at the lower boundary of the gap falls below the light line. The triangular lattice structure has this property for sufficiently large air filling fractions. In figure B10.19, the bandgap diagram for a structure with a filling fraction of 70% is shown [76]. It can be seen that several bandgaps cross the air line. In order to confine an air-guided mode within these bandgaps, a core defect consisting of a rather large air hole must be introduced. In the inset of figure B10.19, a structure is shown in which the core defect has been obtained by replacing the central cladding hole and its six nearest neighbours by a single air hole. This structure was found to support both a fundamental and a second-order guided mode, though not at the same frequency. The traces of the two modes are shown in the upper inset of figure B10.19: inside the bandgap the modes are bound, while outside they appear as leaky resonances within the bands of cladding modes. For both modes, guidance only takes place in a rather narrow frequency interval. Furthermore, the guidance properties of the structure were found to depend strongly on the radius of the core defect. For these reasons, the fabrication of air-guiding fibres for practical applications at specific wavelengths requires a high degree of control over the manufacturing process.

Experimentally, air-guiding fibres were realized by Cregan *et al* [12]. In fibres with pitches around 5 μm and air filling fraction of 30–50% several air-guiding transmission bands were observed over a fibre length of a few centimetres. These findings are in good accordance with the theoretical results discussed earlier. The lower air-filling fraction utilized in the experiment was probably sufficient because

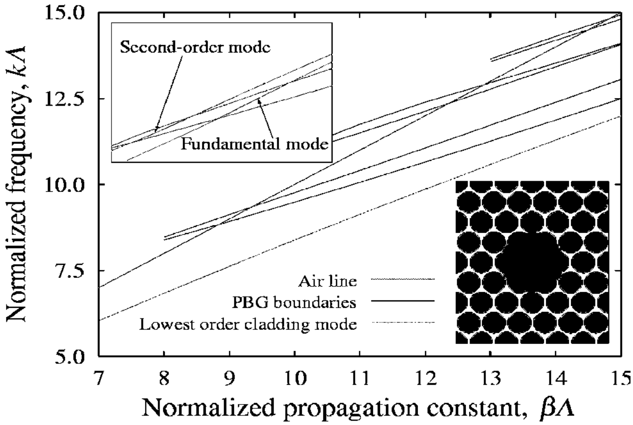


Figure B10.19. PBG boundaries and defect mode traces (upper inset) for the air-guiding PBG fibre discussed in the text. The defect modes are confined when they fall inside the PBG, and leaky elsewhere. The fibre microstructure is shown in the lower inset. Figure from [76].

the fibre had a somewhat modified cladding structure with small interstitial air holes located between the larger ones on the triangular lattice. Due to the relatively short lengths of fibre fabricated, it is difficult to estimate their loss properties, but the detection of guided modes in such structures is an important and encouraging result. A promising recent development was presented by Venkataraman *et al* in [77], when a loss as low as 13 dB km^{-1} was reported for an air-guiding optical fibre. The reported structure is rather close to the kind of fibre structure shown as the inset of figure B10.19, and represents a very significant loss improvement, which indicates a tremendous step towards practical applications of air-guiding PBG fibres. If it becomes possible to manufacture such fibres with even lower losses and better control over the guidance properties, many applications can be envisaged. These include low-loss/high-power transmission at wavelengths where the absorption loss of silica is high, gas-sensing applications, and spectral filtering exploiting the narrowness of the guided-mode transmission bands [12]. Applications for laser-assisted atom transport or even as particle accelerators have also been proposed [78].

B10.6 Conclusions

Microstructured optical fibres have rapidly matured not only as a research field, but also as commercial products. As this paper has demonstrated, index-guiding HFs based on a cladding structure with a triangular lattice of air holes can now routinely be fabricated with useful tolerances over a wide parameter range, and more complicated structures, such as double-clad fibres with active core/inner cladding have also been demonstrated. The unique properties and design flexibility of these fibres open up a wide range of possible applications as functional components in fibre communication networks (e.g. for amplification or dispersion compensation), in broadband sources or for high-power transmission. Such applications of microstructured fibres have already been actively investigated for several years. Other possibilities have up until now only been touched briefly upon. These include applications as gas- or biosensors, and the fabrication of active (for instance, electro-optic) components by filling the fibre holes with some form of active material. Furthermore, the potential of PBG fibres is largely unexplored, in particular from the experimental point of view. As has been discussed here, these fibres have some unique possibilities such as guiding light in an air core, or guiding only a single polarization mode. However, more information on the loss properties and error tolerances of these fibres is needed before definite statements about their usefulness can be made.

One of the objectives of this paper has been to describe a range of recent research results that exploit the unique features of existing HFs. One area in which significant progress has been made recently is the field of highly nonlinear index-guiding HFs. The generation of broad supercontinuum spectra, and all-optical data regeneration, are examples of the significant advances that have resulted from this silica-based HF technology. Moreover, nonsilica index-guiding HFs promise to offer nonlinear fibre devices with unprecedentedly low operating powers (1–10 mW) and short device lengths (0.1–1 m).

As HF technology becomes increasingly mature, numerous applications within the area of optical signal processing, amplification, high-power transmission etc may be expected. The many unique features of this new class of fibres will most likely make them key components not only in future telecommunications systems, but also within areas such as optical sensors and metrology, and beyond.

Acknowledgments

Tanya Monro would like to thank a number of her colleagues at the ORC, University of Southampton, who have made many important contributions to the research described within this paper. In particular, warm thanks to Kentaro Furusawa, Vittoria Finazzi, Joanne Baggett, Periklis Petropoulous, Walter Belardi, Ju Han Lee, Jonathan Price and David Richardson. Tanya Monro also acknowledges the support of a Royal Society University Research Fellowship.

Jesper Lægsgaard and Anders Bjarklev would like to thank all our colleagues at COM, DTU and at Crystal Fibre A/S for the collaboration and results which enter this work. Special thanks should be given to Mr René Engel Kristiansen and Dr Jes Broeng at Crystal Fibre A/S for particularly great help during the writing of this manuscript.

References

[1] Birks T A, Roberts P J, Russell P St J, Atkin D M and Shepherd T J 1995 Full 2-D photonic band gaps in silica/air structures *Electron. Lett.* **31** 1941

[2] Knight J C, Birks T A, Atkin D M and Russell P St J 1996 Pure silica single-mode fibre with hexagonal photonic crystal cladding *Optical Fiber Communication Conference (vol 2)* p CH35901

[3] Monro T M, West Y D, Hewak D W, Broderick N G R and Richardson D J 2000 Chalcogenide holey fibres *Electron. Lett.* **36** 1998–2000

[4] van Eijkelenborg M, Large M, Argyros A, Zagari J, Manos S, Issa N A, Bassett I M, Fleming S C, McPhedran R C, deSterke C M and Nicorovici N A P 2001 Microstructured polymer optical fibre *Opt. Express* **9** 319

[5] Monro T M, Bennett P J, Broderick N G R and Richardson D J 2000 Holey fibers with random cladding distributions *Opt. Lett.* **25** 206–208

[6] Monro T M, Richardson D J, Broderick N G R and Bennett P J 1999 Holey optical fibers: an efficient modal model *J. Lightwave Technol.* **17** 1093–1101

[7] Broderick N G R, Monro T M, Bennett P J and Richardson D J 1999 Nonlinearity in holey optical fibers: measurement and future opportunities *Opt. Lett.* **24** 1395–1397

[8] Knight J C, Birks T A, Cregan R F, Russell P St J and De Sandro J-P 1998 Large mode area photonic crystal fibre *Electron. Lett.* **34** 1347–1348

[9] Birks T A, Knight J C and Russell P St J 1997 Endlessly single-mode photonic crystal fiber *Opt. Lett.* **22** 961–963

[10] Knight J C, Arriaga J, Birks T A, Ortigosa-Blanch A, Wadsworth J W and Russell P S t J 2000 Anomalous Dispersion in Photonic Crystal Fiber *IEEE Photonics Technol. Lett.* **12** 807–809

[11] White T P, McPhedran R C, deSterke C M, Botten L C and Steel M J 2001 Confinement losses in microstructured optical fibers *Opt. Lett.* **26** 1660–1662

[12] Cregan R F, Mangan B J, Knight J C, Birks T A, Russell P St J, Roberts P J and Allan D C 1999 Single-mode photonic band gap guidance of light in air *Science* **285** 1537–1539

[13] Broeng J, Barkou S E, Bjarklev A, Knight J C, Birks T A and Russell P St J 1998 Highly increased photonic band gaps in silica/air structures *Opt. Commun.* **56** 240–244

[14] Broeng J, Søndergaard T, Barkou S, Barbeito P and Bjarklev A 1999 Waveguidance by the photonic bandgap effect in optical fibres *Pure Appl. Opt.* **1** 477–482

[15] Broeng J, Mogilevstev D, Barkou S E and Bjarklev A 1999 Photonic crystal fibers: a new class of optical waveguides *Opt. Fiber Technol.* **5** 305–330

[16] Bjarklev A, Broeng J, Barkou S E and Dridi K 1998 Dispersion properties of photonic crystal fibres *24th European Conference on Optical Communication (ECOC'98, Madrid, vol 1)* pp 135–136.

[17] Riishede J, Libori S B, Bjarklev A, Broeng J and Knudsen E 2001 *Proc. 27th European Conference on Optical Communication (ECOC' 2001)* Th.A1.5

[18] Ho K M, Chan C T and Soukoulis C M 1990 Existence of a photonic gap in periodic dielectric structures *Phys. Rev. Lett.* **65** 3152–3155

[19] Meade R D, Rappe A M, Brommer K D, Joannopoulos J D and Alerhand O L 1993 Accurate theoretical analysis of photonic band-gap materials *Phys. Rev. B* **48** 8434–8437

[20] Eggleton B J, Westbrook P S, Windeler R S, Spalter S and Strasser T A 1999 Grating resonances in air/silica microstructured optical fibers *Opt. Lett.* **24** 1460–1462

[21] Brechet F, Marcou J, Pagnoux D and Roy P 2000 Complete analysis of the characteristics of propagation into photonic crystal fibers, by the finite element method *Opt. Fiber Technol.* **6** 181–191

[22] Mogilevtsev D, Birks T A and Russell P St J 1998 Group-velocity dispersion in photonic crystal fibers *Opt. Lett.* **23** 1662–1664

[23] Monro T M, Richardson D J, Broderick N G R and Bennett P J 2000 Modelling large air fraction holey optical fibers *J. Lightwave Technol.* **18** 50–56

[24] Monro T M, Broderick N G R and Richardson D J 2000 Exploring the optical properties of holey fibres *NATO Summer School on Nanoscale Linear and Nonlinear Optics (Erice, Sicily)*

[25] Bennett P J, Monro T M and Richardson D J 1999 Towards practical holey fibre technology: fabrication splicing modeling and characterization *Opt. Lett.* **24** 1203–1205

[26] Steel M J, White T P, de Sterke C M, McPhedran R C and Botten L C 2001 Symmetry and degeneracy in microstructured optical fibers *Opt. Lett.* **26** 488–490

[27] Poladian L, Issa N A and Monro T M 2002 Fourier decomposition algorithm for leaky modes of fibres with arbitrary geometry *Opt. Express* **10** 449–454

[28] Ranka J K, Windeler R S and Stentz A J 2000 Visible continuum generation in air silica microstructure optical fibers with anomalous dispersion at 800 nm *Opt. Lett.* **25** 25–27

[29] Agrawal G P 1989 *Nonlinear Fiber Optics* (New York: Academic)

[30] Okuno T, Onishi M, Kashiwada T, Ishikawa S and Nishimura M 1999 Silica-based functional fibers with enhanced nonlinearity and their applications *IEEE J. Sel. Top. Quant.* **5** 1385–1391

[31] Belardi W, Lee J H, Furusawa K, Yusoff Z, Petropoulos P, Ibsen M, Monro T M and Richardson D J 2002 A 10Gbit/s tuneable wavelength converter based on four-wave mixing in highly nonlinear holey fibre *Proc. ECOC, Copenhagen, Denmark, September 2002* Postdeadline paper PD1.2

[32] Finazzi V, Monro T M and Richardson D J 2002 Confinement loss in highly nonlinear holey optical fibers *Proc. OFC*, OSA *Technical Digest (Anaheim, California, 2002)* pp 524–525

[33] Price J H V, Furusawa K, Monro T M, Lefort L and Richardson D J 2002 Tunable femtosecond pulse source operating in the range 1.06–1.33 micron based on an Yb^{3+}-doped holey fiber amplifier *J. Opt. Soc. Am. B* **19** 1286–1294

[34] Birks T A, Mogilevtsev D, Knight J C and Russell P St J 1999 Dispersion compensation using single material fibers *IEEE Photonics Technol. Lett.* **11** 674–676

[35] Nakazawa N, Kubota H and Tamura K 1999 Random evolution and coherence degradation of a high-order optical soliton train in the presence of noise *Opt. Lett.* **24** 318–320

[36] Petropoulos P, Monro T M, Belardi W, Furusawa K, Lee J H and Richardson D J 2001 2R-regenerative all-optical switch based on a highly nonlinear holey fiber *Opt. Lett.* **26** 1233–1235

[37] Futami F, Watanabe S and Chikama T 2000 Simultaneous recovery of 20 × 20 GHz WDM optical clock using supercontinuum in a nonlinear fiber *Proc. ECOC*

[38] Hansryd J and Andrekson P A 2001 Broad-band continuous-wave-pumped fiber optical parametric amplifier with 49-dB gain and wavelength-conversion efficiency *Photonics Technol. Lett.* **13** 194–196

[39] Druon F, Sanner N, Lucas-Leclin G, Georges P, Gaumé R, Viana B, Hansen K P and Petersson A 2002 Self-compression of 1-um femtosecond pulses in a photonic crystal fiber *Proc. CLEO*

[40] Lee J H, Yusoff Z, Belardi W, Ibsen M, Monro T M, Thomsen B and Richardson D J 2002 A holey fiber based WDM wavelength converter incorporating an apodized fiber Bragg grating filter *Proc. CLEO*

[41] Sharping J E, Fiorentino M, Kumar P and Windeler R S 2002 All optical switching based on cross-phase modulation in microstructure fiber *Photonics Technol. Lett.* **14** 77

[42] Takara H, Ohara T, Mori K, Sato K, Yamada E, Jinguji K, Inoue Y, Shibata T, Morioka T and Sato K-I 2000 Over 1000 channel optical frequency chain generation from a single supercontinuum source with 12.5 GHz channel spacing for DWDM and frequency standards *Proc. ECOC*

[43] Hansen K P, Jensen J R, Jacobsen C, Simonsen H R, Broeng J, Skovgaard P M W, Petersson A and Bjarklev A 2002 Highly nonlinear photonic crystal fiber with zero-dispersion at 1.55 μm *OFC 02 Post deadline*

[44] Farr L, Knight J C, Mangan B J and Roberts P J 2002 Low loss photonic crystal fibre *Proc ECOC, Paper PD1.3 (Copenhagen, Denmark 2002)*

[45] Lee J H, Teh P C, Yusoff Z, Ibsen M, Belardi W, Monro T M and Richardson D J 2002 A holey fiber-based nonlinear thresholding device for optical CDMA receiver performance enhancement *IEEE Photonics Technol. Lett.* **14** 876–878

[46] Yusoff Z, Lee J H, Belardi W, Monro T M, Teh P C and Richardson D J 2002 Raman effects in a highly nonlinear holey fiber: amplification and modulation *Opt. Lett.* **27** 424–426

[47] Nilsson J, Selvas R, Belardi W, Lee J H, Yusoff Z, Monro T M, Richardson D J, Park K D, Kim P H and Park N 2002 Continuous-wave pumped holey fiber Raman laser *Proc. OFC, OSA Technical Digest (Anaheim, California)*

[48] Lee J H, Yusoff Z, Belardi W, Ibsen M, Monro T M, Thomsen B and Richardson D J 2002 A holey fiber based WDM wavelength converter incorporating an apodized fiber Bragg grating filter *CLEO/QELS 2002 (Long Beach, California)*

[49] Coen S, Chau A H L, Leonhardt R and Harvey J D 2002 Supercontinuum generation via stimulated Raman scattering and parametric four-wave mixing in photonic crystal fibers *J. Opt. Soc. Am. B* **19** 753–764

[50] Hansen K P, Larsen J J, Jensen J R, Keiding S, Broeng J, Simonsen H R and Bjarklev A 2001 Super continuum generation at 800 nm in highly nonlinear photonic crystal fibers with normal dispersion *Proc. LEOS*

[51] Hartl I, Li X D, Chudoba C, Ghanta R K, Ko T H and Fujimoto J G 2001 Ultrahigh-resolution optical coherence tomography using continuum generation in an air–silica microstructure optical fiber *Opt. Lett.* **26** 608–610

[52] Drullinger R E, Diddams S A, Vogel K R, Oates C W, Curtis E A, Lee W D, Itano W M, Hollberg L and Bergquist J C 2001 All-optical atomic clocks *International Frequency Control Symposium and PDA Exhibition* pp 69–75

[53] Champert P A, Popov S V and Taylor J R 2002 Generation of multiwatt, broadband continua in holey fibers *Opt. Lett.* **27** 122–124

[54] Husakou A V and Herrmann J 2001 Supercontinuum generation of higher-order solitons by fission in photonic crystal fibers *Phys. Rev. Lett.* **87** 203901-1–203901-4

[55] Herrmann J, Griebner U, Zhavoronkov N, Husakou A, Nickel D, Knight J C, Wadsworth W J, Russell P St J and Korn G 2002 Experimental evidence for supercontinuum generation by fission of higher-order solitons in photonic fibers *Phys. Rev. Lett.* **88** 173901-1–173901-4

[56] Hansen K P, Jensen J R, Birkedal D, Hvam J M and Bjarklev A 2002 Pumping wavelength dependence of super continuum generation in photonic crystal fibers *Proc. Conference on Optical Fiber Communication Proc. OFC*

[57] Ferrando A, Zacarés M, Fernández de Córdoba P and Binosi D 2002 Self-trapped localized modes in photonic crystal fibers *Nonlinear Optics (NLO) Maui*, *Hawaii*, *July 29–Aug 2*

[58] Asobe M 1997 Nonlinear optical properties of chalcogenide glass fibers and their application to all-optical switching *Opt. Fiber Technol.* **3** 142–148

[59] Friberg S R and Smith P W 1987 Nonlinear optical-glasses for ultrafast optical switches *IEEE J. Quantum Electron.* **23** 2089–2094

[60] Monro T M, Kiang K M, Lee J H, Frampton K, Yusoff Z, Moore R, Tucknott J, Hewak D W, Rutt H N and Richardson D J 2002 Highly nonlinear extruded single-mode holey optical fibers *Proc. OFC, OSA Technical Digest 315–317 (Anaheim, California)*

[61] Gambling A *et al* 1978 Measurement of radiation loss in curved single-mode fibres *Microwaves Opt. Acoust.* **2** 134–140

[62] Sakai J I and Kimura T 1978 Bending loss of propagation modes in arbitrary-index profile fibers *Appl. Opt.* **17** 1499–1506

[63] Knudsen E, Bjarklev A, Broeng J and Barkou S E 2000 Macro-bending loss estimation for air-guiding photonic crystal fibres *14th International Conference on Optical Fiber Sensors OFS2000* pp 904–907

[64] Sørensen T, Broeng J, Bjarklev A, Knudsen E, Barkou S E, Simonsen H R and Riis Jensen J 2001 Macrobending loss properties of photonic crystal fibres with different air filling fractions *Proc. ECOC'2001*, *Amsterdam*, *The Netherlands*

[65] Baggett J C, Monro T M, Furusawa K and Richardson D J 2001 Comparative study of large mode holey and conventional fibers *Opt. Lett.* **26** 1045–1047

[66] Baggett J C, Monro T M, Furusawa K and Richardson D J 2002 Distinguishing transition and pure bend losses in holey fibers *Paper CMJ6 CLEO 2002 (Long Beach, California)*

[67] Marcuse D 1982 Influence of curvature on the losses of doubly clad fibres *Appl. Opt.* **21** 4208–4213

[68] Sahu J K, Renaud C C, Furusawa K, Selvas R, Alvarez-Chavez J A, Richardson D J and Nilsson J 2001 Jacketed air-clad cladding pumped ytterbium-doped fibre laser with wide tuning range *Electron. Lett.* **37** 1116–1117

[69] DiGiovanni D J and Windeler R S 1999 Article comprising an air-clad optical fiber *United States Patent US* 5 907 652, May 25, 1999

[70] Furusawa K, Malinowski A N, Price J H V, Monro T M, Sahu J K, Nilsson J and Richardson D J 2001 A cladding pumped ytterbium-doped fiber laser with holey inner and outer cladding *Opt. Express* **9** 714–720

[71] Knight J C, Broeng J, Birks T A and Russell P St J 1998 Photonic band gap guidance in optical fibers *Science* **282** 1476–1478

[72] Barkou S E, Broeng J and Bjarklev A 1999 Silica–air photonic crystal fiber design that permits waveguiding by a true photonic bandgap effect *Opt. Lett.* **24** 46–48

[73] Barkou S E, Broeng J and Bjarklev A 1999 Dispersion properties of photonic bandgap guiding fibers *Optical Fiber Communication Conference, Paper FG5, 117–119*, *San Diego*, *Feb. 1999*

[74] Bjarklev A, Broeng J, Barkou S E, Knudsen E, Søndergaard T, Berg T W and Dyndgaard M G 2000 Polarization properties of honeycomb-structured photonic bandgap fibres *J. Opt. A* **2** 584–588

[75] Broeng J 1999 *PhD Thesis* Technical University of Denmark

[76] Broeng J, Barkou S E, Søndergaard T and Bjarklev A 2000 Analysis of air-guiding photonic bandgap fibers *Opt. Lett.* **25** 96–98

[77] Venkataraman N, Gallagher M T, Smith C M, Müller D, West J A, Koch K W and Fajardo J C 2002 Low loss (13 dB/km) air core photonic band-gap fibre *ECOC'2002 Copenhagen, Denmark, Post deadline paper PD1.1*

[78] Xintian E L 2001 Photonic band gap fiber accelerator *Phys. Rev. Special Top.—Accelerators Beams* **4** 051301

B11

Engineered optical materials

Peter G R Smith

B11.1 Introduction

The purpose of this chapter is to review recent developments in engineered optical materials, and in particular nonlinear crystals. This definition can be taken to mean conventional materials that are structurally altered to give new and enhanced optical properties. To distinguish such materials from other structured materials such as doped semiconductors or countless other examples, we will also limit the definition to mean only those that are noncentrosymmetric. Thus it includes periodically poled ferroelectrics and polar crystals. The materials covered in this chapter are ones that were mostly developed for their nonlinear optical properties, but are now seeing applications in other areas of optics such as electro-optics. This review is written for the general laser scientist who wishes to learn more about what can be achieved with these materials, and not for those specialists already working in quasi-phase matched (QPM) materials.

B11.1.1 Overview

To date, the most widely used engineered nonlinear optical material is periodically poled lithium niobate (PPLN). Figure B11.1 shows the number of journal publications per annum over the last 10 years on PPLN and closely related materials such as periodically poled lithium tantalate (PPLT) and doped variants of PPLN. It is clear that there has been a dramatic development of this technology in the last decade. The original concept of QPM dates back to the paper by Armstrong [1], after which various techniques such as periodic poling during crystal growth, and ion diffusion to induce domain inversion were used. However, the great upsurge in interest came following the successful demonstration of pulsed electrical field poling by Yamada [2]. The majority of these publications have made use of the quasi-phase matching property of domain reversal gratings to allow efficient nonlinear optical interactions. This quasi-phase matching technique overcomes the inherent dispersion of materials by periodic reversals of the nonlinear coefficient of the material. A detailed derivation of quasi-phase matching will be presented in section B11.5.

PPLN and its close relatives do not hold a monopoly on quasi-phase matching, and indeed quasi-phase matching can be carried out in materials by periodically modulating the strength of the nonlinearity rather than by inverting the sign of the nonlinearity. This technique is particularly attractive in the area of poled glasses where it is often simpler to periodically degrade the nonlinearity by UV or electron beam exposure [3]. The poled glass materials are very promising from a technological perspective because of their compatibility with optical transmission fibre. However, to date, the nonlinearity is significantly lower than that achievable in ferroelectrics, significantly less than 1 pm/V compared with >10 pm/V, and so applications are limited. Another research area for QPM is in GaAs

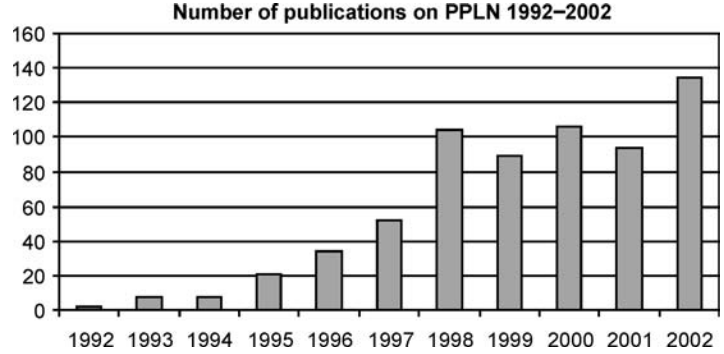


Figure B11.1. Growth in the number of journal publications on PPLN and related materials from 1992 to 2002.

where various techniques have been used to make structurally reversed materials for use in long wavelength generation. Recent techniques include direct bonding of stacked materials [4] and templated growth [5].

Many other ferroelectrics can be used to produce QPM materials, in particular the KTP family materials have a number of advantages, particularly for high energy applications where their higher nanosecond-pulse optical-damage thresholds and smaller temperature coefficients provide advantages over PPLN despite lower nonlinearity. More exotic ferroelectric materials have also seen applications where larger transmission ranges are desirable, and in particular $BaMgF_2$ for UV and SBN for longer wavelengths. The primary ferroelectric materials for QPM are reviewed in table B11.1.

QPM materials find application as bulk crystals for many applications, but equally there are many applications where waveguide devices are more desirable, as they are able to prevent the deleterious effects of diffraction. By tightly confining the optical power within the waveguide core it is possible to achieve both higher absolute conversion efficiencies and lower input powers.

B11.1.2 Overview of applications of QPM materials

Having introduced the most common QPM materials, it is now appropriate to discuss the applications of these materials in optics and elsewhere. The vast majority of applications involve the use of the materials for optical frequency conversion, however, more recently, the properties of domain

Table B11.1. The most common QPM materials and their principle advantages.

Material	Transmission range (μm)	Advantages
$LiNbO_3$	0.38–5.2	Cheap, high nonlinearity, available in large sizes (up to 5 cm)
$LiTaO_3$	0.28–5.5	Better UV transmission, slight smaller nonlinearity than $LiNbO_3$
KTP	0.35–3.3	Better power handling in the visible and for nanosecond pulses; smaller temperature coefficients than $LiNbO_3$ or $LiTaO_3$, relatively more expensive

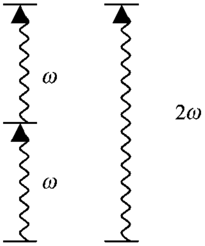
engineering have started to receive attention for other optical applications. In particular, researchers have begun to exploit other properties of domain engineered materials such as their differential etching, electro-optic or pyroelectric properties. These novel applications will be reviewed later in section B11.8.

The desirable nonlinear optical properties of QPM materials have provided most of the impetus behind their development in recent years. They have the enormous advantage of freeing nonlinear device performance from the limitations resulting from the naturally available birefringence of conventional crystals. Subject only to the proviso that the structure must permit fabrication, QPM materials allow efficient nonlinear optical interactions throughout the whole transmission range of the material. Furthermore, by choosing appropriate grating design, it is possible to tailor the nonlinear response to match the particular geometry, spectrum, temporal and chirp properties, temperature bandwidth and configuration of the nonlinear interaction. This flexibility provides a tremendous opportunity for new and exciting device performance, particularly for short pulse applications.

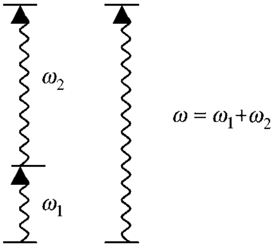
B11.2 Second order nonlinear processes

The majority of nonlinear interactions in QPM materials make use of the second-order nonlinearity $\chi^{(2)}$, which in general mediate the interactions of three interacting waves in a material. The second-order nonlinearity may be used in a number of processes depending on the frequencies of the three waves and on the boundary conditions in the interaction (such as feedback mirrors) and input waves. The most common interaction is second harmonic generation (SHG) in which a new wave is created with half of the input wavelength or equivalently twice the frequency, the most familiar example of which is probably in frequency doubling of the 1064 nm Nd:YAG laser to give visible green at 532 nm. SHG can be seen as a special case of a three wave interaction in which the two input fields are degenerate. In general these $\chi^{(2)}$ processes are called optical parametric interactions, and they provide a rich set of possibilities for applications in optics. Closely related to SHG are the processes of sum and difference frequency generation (SFG and DFG, respectively) in which the output occurs at the sum or difference frequency of two input waves. In addition to SHG, SFG and DFG there also exist optical parametric processes in which an input photon is split to create two output photons at longer wavelengths, subject to the constraint of conversion of energy. This distinction is important; in the SHG, SFG and DFG processes the output frequencies created are set by the input frequencies in contrast to parametric processes in which any energy conserving process is possible (i.e. any pair of photons for which the sum of their energies adds up to the input energy). In a parametric process the phase matching requirements decide which waves see coherent growth. Figure B11.2 shows schematically the various nonlinear mixing processes of SHG, sum frequency generation and generalized optical parametric generation.

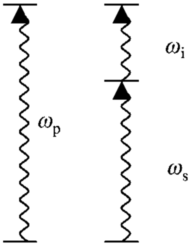
The optical parametric approach can be used to make devices which in many ways mirror the operation of lasers with the parametric process providing a gain medium through parametric amplification rather than the stimulated emission process that occurs in a conventional laser. These devices are called optical parametric oscillators (OPOs) and can provide a widely tunable light output throughout the visible and near IR. They present a valuable enabling tool in spectroscopy and remote sensing. Optical parametric oscillators may be characterized by the feedback schemes used into, singly resonant in which a single output wave is fed back into the cavity, and doubly resonant in which feedback occurs at both output wavelengths. Useful nomenclature is the use of self-explanatory term ‘pump’ for the input laser pump. In addition, the terms ‘signal’ and ‘idler’ are widely used for the two output waves, with ‘signal’ usually referring to the more energetic of the two output waves, although it is occasionally used to refer to the resonated wave, even if it is of longer wavelength. Continuing the similarity with conventional lasers it is possible also to use parametric devices as amplifiers known as



Second harmonic generation



Sum frequency generation



Optical parametric generation

Figure B11.2. Illustration of energy conservation in $\chi^{(2)}$ nonlinear materials.

optical parametric amplifiers (OPA) or at very high gains in which amplification of parametric fluorescence is used as an optical parametric generator (OPG). Figure B11.3 shows a schematic of an optical parametric oscillator.

Optical parametric oscillators provide versatile sources with a key feature being that they are able to produce light at almost any wavelength within the transmission of the material. To produce light efficiently the process must also be phase matched, and thus it is possible to change the wavelength of an OPO by changing angle, QPM period, temperature or pump wavelength. Thus optical parametric oscillators are valuable for converting light from available laser transitions to longer wavelengths at which no suitable laser transitions occur.

B11.3 Materials

Nonlinear optics is concerned with materials in which the response of the material is a nonlinear function of the incident electric fields. It is useful to think of the response of material in terms of

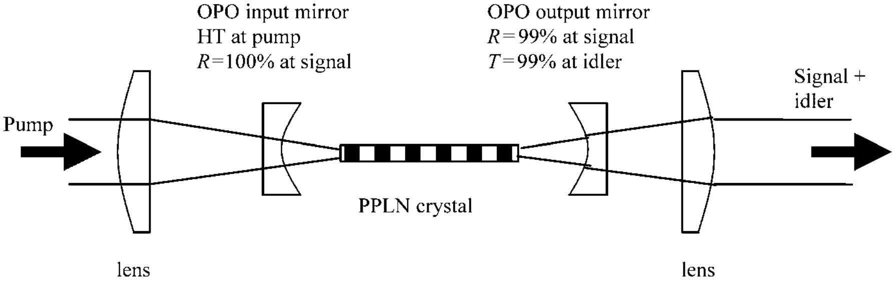


Figure B11.3. Schematic of an optical parametric oscillator.

the polarization set up due to the applied electric fields. In general, the induced polarization will depend on the state of the material (temperature, thermal history, pressure, strain, etc) and also on the frequencies of the applied field(s). The polarization response will normally show a strong dependence on frequency—at low frequencies it will have contributions from orientational, ionic and electronic motion, whereas at higher frequencies (for example, in the visible optical) only the electrons in the material are capable of responding quickly enough to applied optical fields.

It is conventional to describe the polarization in terms of a power series in applied electric field as in equation (B11.1). The polarization can be seen to comprise a number of terms, the first of which is the linear polarization and is responsible for the linear refractive index. The variation of refractive index with wavelength, or equivalently frequency, is known as dispersion, and plays an important role in wavelength conversion.

$$\boldsymbol{P} = \epsilon_0 \left[\chi^{(1)} + \chi^{(2)}\boldsymbol{E}^2 + \chi^{(3)}\boldsymbol{E}^3 + \cdots\right]\boldsymbol{E}. \tag{B11.1}$$

The values of the various nonlinear terms in the susceptibility will depend on the material, and will vary with electronic configuration, density, and particularly crystal structure. In general, the coefficients χ will be tensor quantities as they can relate any of the generated polarization to any of the fields present. However, it is possible to make an important general observation regarding the symmetry of the crystal. If the material is centro-symmetric then it will not have any even-order nonlinearities, so that $\chi^{(2)}$, $\chi^{(4)}$, etc, will be zero. The reason for this is that the nonlinear susceptibility must display the symmetry properties of the crystal medium itself, and in a centrosymmetric crystal the inversion operation ($x \rightarrow -x, y \rightarrow -y, z \rightarrow -z$) leaves the crystal unaltered. The same operation on vector quantities such as the polarization $\boldsymbol{P}$ or $\boldsymbol{E}$ reverses their sign, and so it follows from applying inversion to either side of equation (B11.1) that all even order nonlinear coefficients must be zero.

This requirement that the material be noncentrosymmetric provides the unifying feature of these materials, and it is the possibility of structurally modifying them, for example through periodic reversal that enables their unique potential. All polar materials are noncentrosymmetric, common examples are quartz and gallium arsenide, and are thus able to exhibit second-order nonlinearity. However, in these polar materials it is relatively difficult to create spatially structured or reversed patterns, and so the majority of activities have concentrated on the use of ferroelectric materials such as lithium niobate. Patterned domain inversion in materials such as lithium niobate is essential for the fabrication of engineered optical materials, and can be accomplished in a large number of different ways. The most comprehensive reference on ferroelectrics is probably the book by Lines and Glass [6], which provides a complete discussion of these materials.

B11.4 Ferroelectric materials

It is useful to review some of the most important properties of ferroelectric materials to gain an understanding of their uses as engineered optical materials. The ferroelectric materials are a subset of polar materials, in simple terms being those that possess a spontaneous electric dipole moment within the unit cell that is capable of being inverted by the application of a sufficiently large external electric field. In reality the definition is more complex, and for a full discussion one should consult the standard text by Lines and Glass [6]. As each unit cell in a ferroelectric has a permanent electric dipole moment the whole crystal will have what is known as a spontaneous polarization. In the unit cell of a ferroelectric the mean position of negative charge is displaced from the mean position of positive charge leading to the permanent dipole moment.

Within a ferroelectric crystal a region within which all the unit cells have the same polarity is called a domain. Commercially purchased ferroelectric crystals are normally purchased in a single domain state. In general, ferroelectric crystals will exhibit one or more structural phase transitions between the ferroelectric state and other ferroelectric or paraelectric states. These phase transitions can have a dramatic effect on a crystal, and great care must be exercised when taking a crystal through a phase transition. The temperature at which the material becomes a ferroelectric is known as the Curie temperature. For lithium niobate, for example, the Curie temperature is 1145°C compared to a melting point of 1240°C. A comprehensive treatment of lithium niobate can be found in the book by Prokhorov and Kuz'minov [7].

Another important property of ferroelectrics is pyroelectricity—which is the appearance of charge on certain surfaces of the material as it is heated or cooled. This pyroelectric charge is caused by variation of the spontaneous polarization with temperature. The relative position of different ions will change with temperature, causing a change in dipole moment. In a free crystal the spontaneous polarization is neutralized on the crystal's surface by free charge, and as the spontaneous polarization varies charges appear on the surfaces.

The application of a sufficiently large electric field to a ferroelectric can cause the domain structure to reverse, and thus the mean positions of the positive and negative charges are swapped. This movement of electrical charge constitutes a displacement current, and so an electrical current must be supplied by the external poling circuit. The amount of charge supplied controls the poled area (A), so that $Q = 2P_s A$. For lithium niobate $P_s = 72\,\mu\mathrm{C\,cm^{-2}}$. The subject of domain inversion in ferroelectrics is well researched and can be found in a number of books [6, 8]. However, despite all this research, practical methods of controlling domain inversion remain something of a black art.

In ferroelectric materials, the domain formation is strongly influenced by the underlying crystal symmetry. A widely adopted model for the electric field poling process involves growth of domains from one polar face of the crystal to the other followed by sideways growth of the domain wall. It is widely noticed that the domain formation habit is strongly influenced by the symmetry of the crystal. A striking example comes in $LiNbO_3$ with its 3m symmetry. It can be seen from figure B11.4, which shows an enlargement of a section of PPLN crystal, that the domain walls are predominantly found to lie along the x-direction (vertical in the image) and at ±120° to that axis. So in designing a PPLN grating the sample is usually oriented so that light travels along the y-direction, with the domain walls running along the x-direction. The underlying symmetry properties of the crystal dominate the poling process, and in the image it is clear that at the ends of gratings bars the domains terminate in facets at ± 120° despite the fact that the electrodes that produced the PPLN had squared-off ends! Consequently, it is important that the desired pattern be compatible with the preferred domain habit of the material.

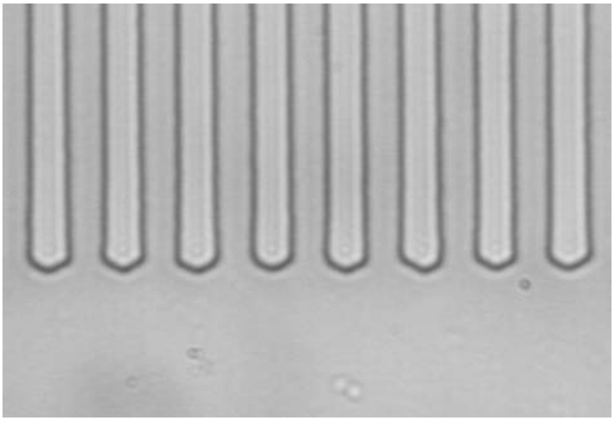


Figure B11.4. Photomicrograph of +z face of PPLN crystal showing domains terminating in facets at ±120° to the *x*-direction (vertical in image).

B11.5 Quasi-phase matched materials

The quasi-phase matching technique has a long history. It was first proposed in 1962 by Armstrong [1], but was largely eclipsed by the development of birefringent phase matching 1962 [9, 10]. QPM continued to attract interest throughout the 1970s, concentrating largely on materials with domains formed during crystal growth. An excellent article covering the history and development of QPM materials by one of the major figures can be found in [11].

The realization that in-diffusion of dopants could be used to reverse the domain direction close to the crystal surface and thus make periodic structures led to a rapid growth in research into QPM waveguides in materials such as $LiNbO_3$ and KTP. However, the critical step came in the work by Yamada [2], in which bulk domain inversion by using a periodic electrode was used for the first time to form PPLN. It is interesting that whilst this process was revolutionary in optics, the idea of forming a domain pattern in a ferroelectric by the use of patterned electrodes and a high-field pulse was well established for ferroelectric memories in the 1970s [8].

The basic idea of QPM is to periodically reverse the direction of the nonlinearity in the material so that the phase of the nonlinearly generated light adds in phase with the light generated earlier in the crystal. The phase mismatch occurs because the different wavelengths in the nonlinear interaction have different refractive indices due to dispersion and thus different phase velocities. As we shall see later, the phase of the nonlinearly generated light depends upon the phase of driving fields and the direction of the nonlinearity, and so periodic reversal of the nonlinearity allows constructive growth of the generated light.

In many ways it is easier to understand these effects via a mathematical derivation of the equations governing the second order interactions. From a personal perspective the clearest exposition of this topic can be found in [12]. The derivation is standard and the starting point is the wave equation which can be manipulated to become:

$$\nabla^2(\boldsymbol{E}) - \mu_0\sigma\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{t}} - \mu_0\epsilon_0\frac{\partial^2 \boldsymbol{E}}{\partial \boldsymbol{t}^2} = \mu_0\frac{\partial^2 \boldsymbol{P}}{\partial \boldsymbol{t}^2}. \tag{B11.2}$$

The right hand side of this equation contains the polarization term, $\boldsymbol{P}$, and it is this term that causes the generation of new frequencies through the susceptibility expansion given by equation (B11.1).

By substituting travelling wave fields in the z-direction, and using slowly varying envelopes for the electric field and polarization, and further by making the slowly varying envelope approximation it is possible to find three coupled equations which describe the second-order optical nonlinearity.

$$\frac{\partial E_{\mathrm{p}}}{\partial z} = \frac{\mathrm{i}\omega_{\mathrm{p}}}{n_{\mathrm{p}}c} dE_{\mathrm{s}}E_{\mathrm{i}}\exp(-\mathrm{i}\Delta kz) \tag{B11.3}$$

$$\frac{\partial E_{\mathrm{s}}}{\partial z} = \frac{\mathrm{i}\omega_{\mathrm{s}}}{n_{\mathrm{s}}c} dE_{\mathrm{p}}E_{\mathrm{i}}^{*}\exp(\mathrm{i}\Delta kz) \tag{B11.4}$$

$$\frac{\partial E_{\mathrm{i}}}{\partial z} = \frac{\mathrm{i}\omega_{\mathrm{i}}}{n_{\mathrm{i}}c} dE_{\mathrm{p}}E_{\mathrm{s}}^{*}\exp(\mathrm{i}\Delta kz). \tag{B11.5}$$

These equations relate the three waves in the material (generally pump, signal and idler E_{p}, E_{s}, E_{i}, respectively). In the equations, n_{α} is the refractive index of wavelength λ_{α}, c is the speed of light, ω_{α} is the angular frequency, and d_{eff} is the nonlinear coefficient. The wavevector mismatch Δk is given by $\Delta k = k_{\mathrm{p}} - k_{\mathrm{s}} - k_{\mathrm{i}}$. As usual, $k_{\alpha} = 2\pi n_{\alpha}/\lambda_{\alpha}$.

These three equations are the governing equations for nonlinear optics, they allow for a tremendous richness of solution, especially if mirrors are used around the nonlinear crystal as in the case of an optical parametric oscillator. These equations can be solved in a variety of situations, and the interested reader is referred to the extensive literature [12–14].

There are some general observations that can be made about these equations and their derivation. The first concerns conservation of energy; in their derivation it is necessary to constrain the frequencies included to satisfy the time dependence of the equations. This constraint $\omega_{\mathrm{p}} = \omega_{\mathrm{s}} + \omega_{\mathrm{i}}$, is equivalent to the conservation of energy in the process.

The second observation is that each equation has a spatial dependence of $\exp(\pm \mathrm{i}\Delta kz)$, where Δk is the phase mismatch. Consideration of any of the above equations shows that the rate of change of that field component depends upon the product of the nonlinearity, the strengths of other two fields, and because of the $\exp(\pm \mathrm{i}\Delta kz)$ term, on a periodic phase factor.

To illustrate how quasi-phase matching works consider a simple second harmonic process. In this case we have two fundamental fields interacting to produce a new field at 2ω. If we relabel the general fields as $E_{(\omega)}$ for the fundamental and $E_{(2\omega)}$ as the second harmonic, and simplify to bring out the essential concepts we end up with the following pair of equations.

$$\frac{\partial E_{(\omega)}}{\partial z} = \mathrm{i}\kappa E_{(2\omega)}E_{(\omega)}^{*}\exp(\mathrm{i}\Delta kz) \tag{B11.6}$$

$$\frac{\partial \mathrm{E}_{(2\omega)}}{\partial z} = i\kappa E_{(\omega)}E_{(\omega)}^{*}\exp(-\mathrm{i}\Delta kz). \tag{B11.7}$$

If the field E_{ω} remains constant (as would be the case for low conversion efficiency and thus low depletion of the fundamental), then it is possible to simply integrate equation (B11.7) to get the field strength $E_{(2\omega)}(L)$ at the end of the crystal (L).

$$E(2\omega)_{L} = \int_{0}^{E} \mathrm{d}E(2\omega) = \mathrm{i}\kappa E^{2}(\omega)\int_{0}^{L} \exp(-\mathrm{i}\Delta kz)\mathrm{d}z. \tag{B11.8}$$

Considering the real and imaginary parts of the integral separately means that the output is given by the integral of sine and cosine over many periods—which has no cumulative contribution.

If significant conversion efficiency is desired it is necessary to find a way to deal with the phase mismatch. In birefringent phase matching different polarizations are used to access appropriate refractive indices in a birefringent material to give zero phase mismatch $\Delta k = 0$.

In the quasi-phase matching approach, the material is periodically altered to modify equation (B11.7) so that the nonlinearity is now a function of z. This simple modification provides the power of the QPM technique.

$$E(2\omega)_L = \int_0^E \mathrm{d}E(2\omega) = \mathrm{i}E^2(\omega)\int_0^L \kappa(z)\exp(-\mathrm{i}\Delta k z)\mathrm{d}z. \tag{B11.9}$$

The nonlinearity is made into a periodic function of z, with a period that matches exp $(\mathrm{i}\Delta k)$. The most common technique is to make $\mathrm{d}(z)$ take positive and negative values by inverting the nonlinearity through periodic poling. This means that the integral over z becomes the integral of a rectified sinusoidal function, which has a cumulative contribution with z.

Figure B11.5 illustrates this effect. Curve (*a*) shows the real part of the LHS of the integrand of equation (B11.8), and curve (*b*) shows the integral with respect to z. It is seen that there is no constructive growth with distance. Curve (*c*) again shows the integrand of equation (B11.8), and curve (*d*) shows an appropriate $\kappa(z)$ which matches the polarity of the generated wave. Curve (*e*) shows the product of curve (*c*) and curve (*d*) which becomes a rectified sine wave, and finally curve (*f*) shows the integral of curve (*e*). Curve (*f*) shows a constructive growth in second harmonic field strength along the crystal. Thus we can clearly see that by reversing the sign of κ with an appropriate period we can get a useful QPM output.

The characteristic length over which constructive addition of the fields occurs is known as the coherence length $l_c = \pi/\Delta k$. As is clear from figure B11.5, the QPM period is $\Lambda = 2l_c = 2\pi/\Delta k$. It is worth mentioning that some authors call Λ the coherence length. To calculate the coherence length for a given material, it is necessary to know the variation of refractive index with wavelength. This is most commonly described by a Sellmeier equation, which provides a power series type expression for $n(\lambda, T)$ as a function of wavelength (and often temperature). An example is the one for $LiNbO_3$ developed from QPM experimental data by Jundt [15], which was used to generate figure B11.6 which shows coherence length versus wavelength for SHG at a temperature of 150°C.

The quasi-phase matching technique does result in a reduced nonlinear coefficient because there is a degree of cancellation within each coherence length, resulting in a $d_{\mathrm{eff}} = (2/\pi)d$. For $LiNbO_3$ this means that $d_{\mathrm{eff}} \approx 16\,\mathrm{pm/V}$ compared to $d_{33} \approx 30\,\mathrm{pm/V}$.

Any periodic modulation of the nonlinearity can be used for quasi-phase matching, the optimal is periodic inversion matching the phase mismatch factor. However, it is also possible to use higher order phase matching (in which the inversion has a period that is an integral multiple of the phase mismatch period) and even a modulation scheme in which the nonlinearity is periodically erased. This latter scheme is commonly employed in poled glasses where inversion is harder to achieve.

The order of phase matching for a given period can be expressed as $\Lambda = 2nl_c$, where n is the order. To be efficient n must clearly be an integer. For a simple 50:50 mark-space ratio grating only odd order quasi-phase matching will be obtained, and in fact this provides a sensitive probe of the quality of a QPM structure. If higher order quasi-phase matching is used then d_{eff} becomes $(2/n\pi)d$.

For each nonlinear process, there will be a bandwidth associated over which an efficient interaction will occur. In general, there will be bandwidths associated with wavelength and temperature (and indeed any other parameter that affects the refractive index). An example of a phase matching curve is shown in figure B11.7, together with a theoretical fit. The PPLN in this example had a period of 6.4 μm and a length of 3.2 mm. The laser was a Nd:YLF operating at 1047 nm. The theoretical shape is a sinc^2 function, which can simply be derived from equation (B11.7). The first zero in the efficiency occurs when

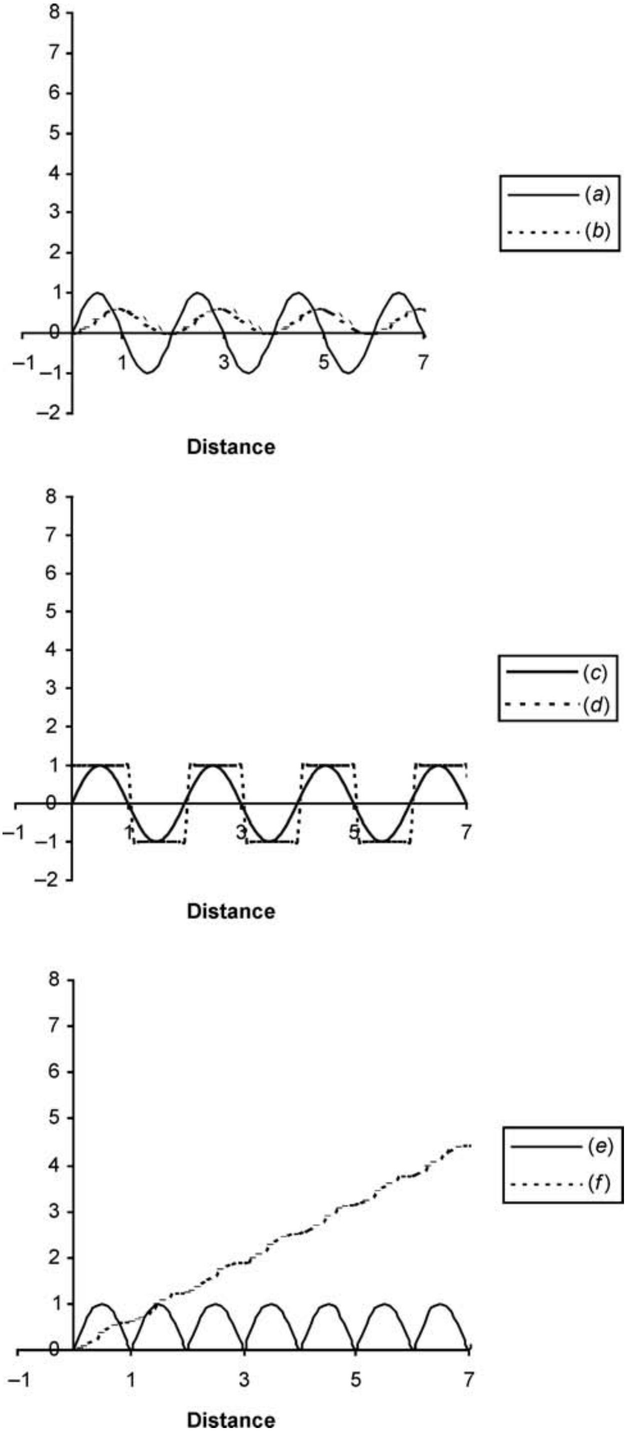


Figure B11.5. Electric field distributions in quasi-phase matching.

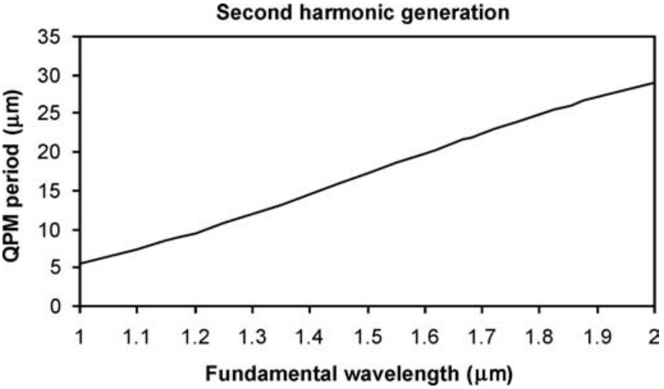


Figure B11.6. The quasi-phase matching period in $LiNbO_3$ for second harmonic generation as a function of fundamental wavelength. The model is calculated at 150°C.

the light generated in the second half of the crystal exactly cancels that generated in the first half. A useful approximate condition for phase matching is that $\Delta kL < \pi/4$. Thus it is clear that the bandwidth is inversely related to the crystal length. General expressions for quasi-phase matching bandwidths can be found in the paper by Fejer [16].

B11.6 Nonlinear processes

The three coupled equations provide a way of characterizing and understanding the various nonlinear processes possible in a nonlinear material, but generally there is little difference between a correctly designed QPM material and a birefringently phase matched material. Consequently, as there are many good discussions of nonlinear optics there is little point in discussing them further here. On a purely personal basis the author would suggest the books on nonlinear optics already mentioned as well as the one by Koechner [17].

All of the standard nonlinear conversion processes have been demonstrated in QPM materials, where their higher nonlinearity and availability in longer lengths makes them a material of choice.

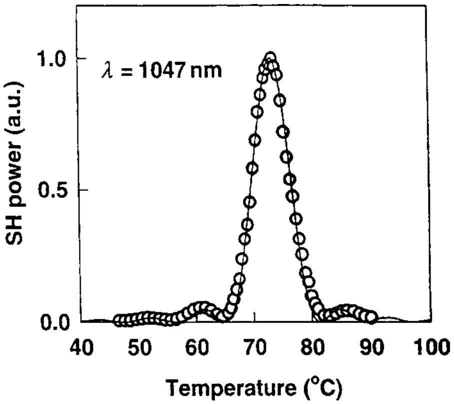


Figure B11.7. Temperature tuning curve showing sinc^2 response for second harmonic generation in PPLN at low conversion efficiency.

In addition, they also provide attractive materials for the observation of other nonlinear processes, such as cascaded nonlinearity and spatial soliton behaviour.

In real world situations it is almost always the case that the maximum conversion efficiency is desired, and this leads to the use of focused beams or the use of waveguides to maximize the efficiency. In this regime, the plane wave treatments do not strictly hold; however, they can still often be applied provided that care is used. Three cases are worth looking at:

(1) The simplest case is a plane wave interaction. In this case, the conversion efficiency scales with length squared. If loosely focused gaussian laser beams are used then the conversion efficiency becomes:

$$\eta = \frac{2\omega^2 d_{\mathrm{eff}}^2 L^2 I(\omega)}{n^3 c^3 \epsilon_0} \operatorname{sinc}^2\left(\frac{\delta k L}{2}\right). \tag{B11.10}$$

This can be applied to an unfocused Gaussian beam such that $I = P/A$ where P is power and the area $A = \pi\omega_0^2/2$. In this equation $\delta k = \Delta k - (2\pi/\Lambda_{\mathrm{grating}})$.

(2) If optimal conversion is desired, then it is necessary to move to tighter focusing. This case received extensive attention by Boyd and Kleinman [18]. In this case there is an issue that focusing to a tighter spot size will result in a higher power at the beam waist, but that this tight focusing will cause the beam to diverge more strongly, and thus become larger at the front and rear of the crystal. The analysis by Boyd and Kleinman shows that there is an optimal focused spot size for a given length of crystal. In this case it is found that the conversion efficiency scales as L rather than L^2. Thus in this case the conversion efficiency takes on units of %/W cm, and a useful (very approximate) number to remember is that for PPLN this number becomes of order 1%/W cm (obviously depending on the wavelength). Thus with a 1 cm crystal, and say 200 mW of input power the efficiency will be of order 0.2% = 0.4 mW. If the crystal is 4 cm long, and the input power is 400 mW then this increases to 1.6% = 6.4 mW. A full calculation for PPLN will lead to a slightly higher number, and depends quite critically on the details of the laser source. It is important to stress that these example numbers are 'ball-park', and act only as a guide…

Figure B11.8 shows a conversion efficiency curve for frequency doubling of a 946 nm Nd:YAG laser. The experiment used optimum focusing and generated up to 450 mW of blue light at 473 nm. To achieve this result, the fundamental laser was used in a relaxation oscillation mode to increase the peak power and hence conversion efficiency.

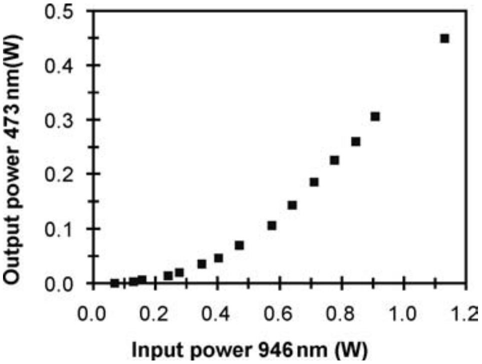


Figure B11.8. Blue light second harmonic generation.

(3) The third case is the waveguide in which the optical confinement provided by the waveguiding from the core compensates the diffraction of the beam. Consequently, a higher intensity results, but at the expense of greater complication in fabrication. The expression for conversion efficiency now becomes more complex, and a calculation must take into account the spatial overlap integral between the nonlinearity and the guided modes. In this case the efficiency becomes once more proportional to L^2, and will be seen quoted as %/W cm^2. In PPLN waveguides, this efficiency can be of order 100%/W cm^2, although in this case launched powers are likely to be far smaller. However the higher efficiency more than compensates, and in a proton-exchange PPLN waveguide 99% pump depletion has been demonstrated by Parameswaran [19].

In conclusion, unfocused beams result in a simple expression for the conversion efficiency, and this regime is often used to characterize material. The optimal focusing regime is ideal for maximizing conversion efficiency of high power lasers, but is subject to not exceeding the damage threshold of the material! Waveguides are ideal for low power conversion but are considerably more complicated to fabricate and their analysis is more complex.

B11.7 Periodically poled lithium niobate

The choice of periodically poled material for a given application depends on a number of factors; most commonly, availability of starting material, cost of starting material, ease of periodic poling, optical transmission, obtainable nonlinearity, damage threshold, short-pulse walk-off, etc. The order here is deliberate, the first two factors tend to dominate which materials are routinely used, and indeed one of the great advantages of periodic poling is their potential for engineering to fit an application. The great success of lithium niobate as a QPM material comes about to a large extent because of its low cost. As an approximate example, a 76 mm diameter lithium niobate wafer costs around \$100–200, whereas potassium titanyl phosphate (KTP) will cost around \$100 for a 10–15 mm square piece. Thus the cost per area for KTP is around 50 times higher than for lithium niobate. While the exact multiplier can be debated, there is no doubt that the fact that lithium niobate is produced in quantities of several tonnes per annum for the surface acoustic wave industry has been one of the major factors in its dominance.

PPLN has proved extremely versatile. It has been successfully operated from the blue [2, 20, 21], through to the mid-IR [22]. It can be fabricated into gratings with periods from a few microns upwards, and in lengths of up to 5 cm. PPLN samples are usually fabricated by the electrical poling technique, where the coercive field is around 22 kV mm^{-1}. Most workers place a large ballast resistor in series with the crystal to limit the current flowing in the sample, and then apply a high voltage in excess of the coercive field. By limiting the time for which the field is applied, it is possible to control the total charge, and thus poled area. An alternative approach, is to control the current in the sample by varying the voltage; this can lead to higher yield and allows excellent control of the poled area.

PPLN has been fabricated in thicknesses of up to 1 mm, but is more commonly available in 500 μm thickness. This limitation on aperture is not normally too restrictive for optical access, but for nanosecond pulses in which the damage threshold becomes energy dependent (around 2 J cm^{-2}) it can be a problem and workers have adopted the approach of bonding samples together to increase the working aperture. In longer pulse regimes and for cw lasers the damage threshold depends more on the intensity, a reasonable figure for which is 150 MW cm^{-2}. In very short pulse applications the damage threshold is often found to depend most strongly on the average power through the crystal, and intensities as high as 2 GW cm^{-2} have been used without damage.

Lithium niobate has a high refractive index ($n_o = 2.2322$, $n_e = 2.1560$ at 1064 nm), leading to strong Fresnel reflections (approximately 14% per surface), and so it is often anti-reflection coated. In OPO applications, this process can actually be more expensive than the crystal itself as complex coatings with multiple transparency bands are needed. In an actual experiment, the PPLN sample is usually mounted in an oven with feedback temperature control. PPLN has relatively large temperature coefficients, which can either be viewed as conveniently useful for tuning of OPO systems, or inconveniently expensive for frequency doubling applications. It is normal to run a PPLN sample at a temperature in excess of 100°C, and often around 150°C to prevent the build-up of photorefractive damage. This effect is caused by charge migration within the crystal, which causes index modulation via the electro-optic effect. The effect is reduced in PPLN relative to unpoled lithium niobate [23], but still remains a major barrier to its application in visible light generation. Another limitation to high power visible operation is the so-called green induced infra-red absorption effect (GRIIRA) [24] that is seen in high power harmonic generation experiments into the visible. In this effect, the green light generated in doubling causes an increase in the infra-red absorption at 1 μm, resulting in heating and eventually in catastrophic damage to the crystal. This limits the visible power to <1 W in the green in PPLN.

The procedure for periodic poling is generally as follows:

(1) selection of appropriate wafers;

(2) cleaning;

(3) definition of electrodes;

(4) electrical poling;

(5) visual characterization;

(6) cutting and polishing;

(7) coating.

The definition of the electrodes is usually done using photolithographic patterning, a mask is used to fabricate a pattern in photoresist which is then used to form the electrode structure. A number of different electrode types can be used, ranging from simple metal electrodes deposited over photoresist, through bar electrodes underneath resist, to contact pressed metal electrodes. It appears possible to get good quality poling with any of these processes, and it is often simpler to use a liquid or gel electrode, although intriguingly the patterned electrode must be placed on the $-z$ face with liquid electrodes whereas the $+z$ face is normally patterned for metal electrodes. The poling process itself remains something of an art, with different groups adopting varying approaches to PPLN fabrication. Figure B11.9 shows schematics of the various steps in the poling process.

PPLN samples are often designed to have multiple gratings with different periods running in parallel tracks along the crystal. This allows tuning of the phase matching by lateral translation of the crystal within the optical beam. An extension of this idea is the use of fanned gratings, in which continuous variation of the period can be achieved. PPLN crystals are almost always housed in small ovens to control the temperature to within a fraction of a degree. The temperature is then optimized to give maximum conversion efficiency for harmonic generation, or is deliberately altered to tune the output wavelengths in an OPO.

B11.7.1 Bandwidth and short pulse operation

The bandwidth of a given grating depends upon the wavelengths used in the interaction and the length of the device [16], it is possible to artificially increase the bandwidth of a grating by deliberately chirping the grating. Such chirped gratings can be advantageously used in recompressing chirped pulses [25]. When dealing with short-pulses it is important to consider the variation in group velocity within the crystal at

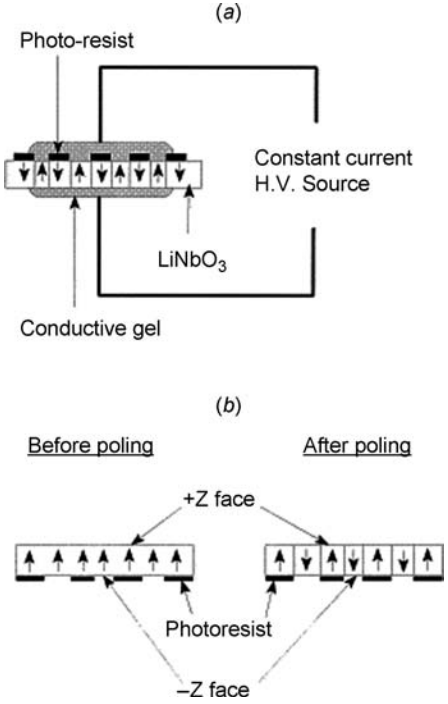


Figure B11.9. Schematic of poling method using conductive gel electrodes patterned at −z face of sample.

different wavelengths, and also the spectral acceptance bandwidth of the crystal for the bandwidth of the pulses. These effects can prove advantageous, and can, for example, provide a method for compressing pulses within an OPO [26]. In short pulse applications the temporal walk-off and bandwidth effects can provide the limit on crystal length, and so when considering different materials it is important to use an appropriate figure of merit to compare the different materials. The figure of merit must be related carefully to the problem in question, as it may often be the case that a material with a lower nonlinear coefficient but lower temporal walk-off of pulses will give better performance.

For many applications variants of lithium niobate are attractive, and there has been considerable recent activity in producing periodically poled materials with superior properties. One of the major drawbacks of PPLN is the photorefractive damage that occurs in visible operation, and to counter this a number of groups have worked on producing magnesium doped materials. Another important recent development has been the use of stoichiometric lithium niobate for periodic poling. Unlike conventional congruent lithium niobate, which has a deficiency of lithium ions from its growth, this material has close to ideal stoichiometry—which results in enhanced nonlinear optical properties and a much lower coercive field. With stoichiometric material it has been possible to pole samples with much larger apertures allowing high power operation.

Another QPM material, which has similar properties to lithium niobate, is lithium tantalate. It is also produced on a large scale for the surface acoustic wave industry, and thus provides an attractive alternative to PPLN. PPLT benefits from better transparency in the near UV (down to 280 nm compared to 380 nm for PPLN). It does, however, have a smaller d_{33} coefficient, and is somewhat more expensive than lithium niobate, so it tends to receive less use.

The other major family of QPM materials is that of KTP. In this material family it is possible to substitute rubidium for potassium and an arsenate group for the phosphate, to produce RTA, and also RTP and KTA. KTP has many advantages, such as higher damage thresholds, and lower temperature coefficients, but against this must be offset the higher price and smaller set of growers for the material. The other major problem seen in the KTP family is the higher ionic conductivity, which can make electrical poling problematic. This can be addressed by several routes, including the use of rubidium ion exchange into the top surface of the crystal [27] and low temperature poling [28] to counter the higher conductivity, or by the use of the hydrothermal growth process which naturally results in lower conductivity crystals. Recent efforts to overcome these problems in flux grown KTP have made use of dopants such as rubidium to control the conductivity [29].

It is hard to compare PPKTP and PPLN as QPM materials, each has advantages and disadvantages, and the correct choice depends on the application, and is often quite subtle.

B11.8 Applications of PPLN and related materials

The number of papers published on PPLN in recent years provides a clear demonstration of the excitement and potential of PPLN and related materials. These applications span a number of industries, wavelength ranges and temporal configurations. These materials can be seen to provide a way of spanning the gaps in the visible and near-IR spectrum at which regular lasers cannot operate, and to provide for applications that need wide tunability. In an even more general sense, these nonlinear materials provide an optical equivalent for the frequency mixing functions so widely used in RF and microwave electronics.

This review is organized into broad industrial areas, and will not pretend to be exhaustive, but is intended to whet the appetite of potential end-users of QPM materials.

B11.8.1 Aerospace

This area is probably the most heavily researched to date, with military applications leading the way. Current military systems tend to operate using 1.064 μm Nd:YAG lasers, and there is a general desire to provide laser systems for longer wavelengths. Firstly to make filtering or observation more difficult, also to exploit the better transmission characteristics of the atmosphere in longer wavelength bands, and also simply to benefit from the better eye-safety obtained at wavelengths such as 1.5 μm.

Remote sensing

This area provides one of the great opportunities for QPM materials, by providing sources tunable to specific wavelength for the detection of trace gases, or atmospheric scattering. Optical parametric oscillators provide flexible and versatile sources for these applications. It is possible to use even simpler systems, such as difference frequency generation, for making low cost and power consumption sources.

Telecomms

This has become a very hot topic for QPM materials in the last few years. The nonlinear conversion of light allows users to provide a wide range of optical processing functions. Applications that are currently receiving attention include wavelength conversion for WDM systems, mid-span spectral inversion, nonlinear dispersion compensation and temporal pulse manipulation for all-optical switching and TDM applications. As these applications need to work with low optical powers they tend to use a waveguide format which utilizes either proton exchange or titanium diffusion to make the waveguides [30–32]. To date, it seems that the variants on proton exchange (soft proton exchange and reverse proton exchange) have provided the best waveguide devices, but this is clearly such an important area that many future developments will occur.

Displays

Although to date the success in this area has been limited, it remains one with great potential. Laser based light shows and displays are already one of the major 'public' applications of the laser. QPM materials can play an important role in helping to provide red, green and blue sources for laser projector systems. A number of schemes are possible, ranging from high power optical parametric oscillators with suitable powers through to fibre laser based systems. The requirements on the materials are stringent, needing several watts of each colour at optimal wavelengths. To date, problems such as GRIIRA in PPLN and grey tracking in KTP provide barriers to the commercial use of the existing QPM materials, but this can be expected to change in the future.

Storage and blue lasers

Historically the need for blue sources for optical disc technology provided a strong impetus for QPM material development; but, with the recent development of blue semiconductor lasers it now seem less likely that QPM materials will find use for read-out lasers. However, it may well be the case that for higher power applications, such as DVD writing and for laser marking, where 10–100 mW is attractive that QPM blue and near-UV sources may become commercial. Further interest in this area comes from the recent increases in fibre laser power such that doubling, tripling, and even frequency quadrupling in QPM materials could provide a viable alternative to gas laser in the blue and near UV. The recent publication of the book by Risk [33] provides an excellent coverage of the applications of blue lasers and a comparison of QPM sources to other types of laser.

Spectroscopy and scientific applications

This provides a tremendous opportunity for QPM materials both as a platform for investigating nonlinear optics, and for creating light sources of unparalleled versatility. Although many conventional nonlinear materials can be used for making optical parametric oscillators, it is likely that higher nonlinearities of periodically poled materials will see them becoming increasingly important in commercial systems. In addition to their use as IR sources for spectroscopy, there are a host of applications in quantum optics experiments where the ability tailor the nonlinearity is likely to lead to new devices and experiments. This trend particularly applies in periodically poled waveguides and poled glass fibres where the high degree confinement leads to high conversion processes and efficient capture of generated light [34].

Frequency conversion does not provide the only application areas for domain engineered materials. Particularly interesting are applications in modulators and beam deflectors where domain engineering can allow new functionality and confer advantages. One example comes in waveguide lithium niobate modulators, in which it has been shown that domain reversal can be used to provide phase matching between the RF drive and the optical wave without having to use segmented electrodes [35]. Domain engineering also provides ways of making new devices for bulk beam modulation, through the creation of electro-optic Bragg gratings [36]. Closely related, although not yet demonstrated would be the development of very fine period domain gratings within waveguides to make switchable retro-reflective Bragg gratings, although as these would require a period of a fraction of a micron so they are perhaps not likely to be realized in any time soon.

Another class of domain engineered devices which do not require gratings are the beam-deflector devices based on refraction [37]. These devices are based on prisms formed from poled electro-optic material and rely on the fact that the change in refractive index in the up and down regions is of opposite sign. Thus by passing a beam through one or more interfaces the beam will be deflected by Snell's law allowing the creation of very fast all solid-state deflectors and scanners.

B11.9 Future of QPM materials

It should be clear from this discussion that QPM materials provide one of the most exciting and vibrant areas of optical research, and it is likely that they will gradually come to replace many of the uses of birefringently phase matched materials. The advantages of QPM are compelling, but clearly many challenges remain. In particular, it is clear that all of the existing QPM materials have drawbacks when used in certain ways, and it is likely that new materials will be developed, either as variants of existing materials (such as the recent works on stoichiometric lithium niobate) or in new ways of making composite materials (QPM GaAs), or even in wholly new nonlinear materials. Another important area of QPM research will be the development of more sophisticated grating designs optimized for high efficiency and specific spectral and temporal response.

References

[1] Armstrong J A, Bloembergen N, Ducuing J and Pershan P S 1962 Interactions between light waves in a nonlinear dielectric *Phys. Rev.* **127** 1918–1939

[2] Yamada M, Nada N, Saitoh M and Watanabe K 1993 First-order quasiphase-matched $LiNbO_3$ waveguide periodically poled by applying an external field for efficient blue second-harmonic generation *Appl. Phys. Lett.* **62** 435–436

[3] Kazansky P G and Pruneri V 1997 Electric-field poling of quasi-phase-matched optical fibers *J. Opt. Soc. Am. B* **14** 3170–3179

[4] Gordon L, Woods G L, Eckardt R C, Route R R, Feigelson R S, Fejer M M and Byer R L 1993 Diffusion-bonded stacked GaAs for quasi-phase-matched 2nd-harmonic generation of a carbon-dioxide laser *Electron. Lett.* **29** 1942–1944

[5] Ebert C B, Eyres L A, Fejer M M and Harris J S 1999 MBE growth of antiphase GaAs films using GaAs/Ge/GaAs heteroepitaxy *J. Cryst. Growth* **201** 187–193

[6] Lines M E and Glass A M 1977 *Principles and Applications of Ferroelectrics and Related Materials* (Oxford: Oxford University Press)

[7] Prokhorov A M and Kuz'minov Yu S 1990 *Physics and Chemistry of Crystalline Lithium Niobate* (Bristol: Institute of Physics)

[8] Burfoot J C and Taylor G W 1979 *Polar Dielectrics and Their Applications* (London: Macmillan)

[9] Giordmaine J A 1962 Mixing of light beams in crystals *Phys. Rev. Lett.* **8** 19–20

[10] Maker P D, Terhune R W, Nisenoff M and Savage C M 1962 Effects of dispersion and focusing on the production of optical harmonics *Phys. Rev. Lett.* **8** 21–22

[11] Byer R L 1997 Quasi-phasematched nonlinear interactions and devices *J. Nonlinear Optical Phys. Mater.* **6** 549–592

[12] Byer R L 1977 Chapter 2—Parametric oscillators and nonlinear materials *Nonlinear Optics*, eds P G Harper and B S Wherrett (London: Academic Press) p 47–160

[13] Shen Y R 2002 *The Principles of Nonlinear Optics* reprint edition (New York: Wiley)

[14] Yariv A 1989 *Quantum Electronics* 3rd edn (New York: Wiley)

[15] Jundt D H 1997 Temperature-dependent Sellmeier equation for the index of refraction n_e in congruent lithium niobate *Opt. Lett.* **22** 1553–1555

[16] Fejer M M, Magel G A, Jundt D H and Byer R L 1992 Quasi-phasematched second harmonic generation: tuning and tolerances *IEEE J. Quantum. Electron.* **28** 2631–2654

[17] Koechner W 1999 *Solid State Laser Engineering* (New York: Springer)

[18] Boyd G D and Kleinman D A 1968 Parametric interaction of focused gaussian light beams *J. Appl. Phys.* **19** 3597–3639

[19] Parameswaran K R, Kurz J R, Roussev R V and Fejer M M 2002 Observation of 99% pump depletion in single-pass second-harmonic generation in a periodically poled lithium niobate waveguide *Opt. Lett.* **27** 43–45

[20] Ross G W, Pollnau M, Smith P G R, Clarkson W A, Britton P E and Hanna D C 1998 Generation of high-power blue light in periodically poled $LiNbO_3$ *Opt. Lett.* **23** 171–173

[21] Yamamoto K, Mizuuchi K, Kitaoka Y and Kato M 1995 Highly efficient quasi-phase-matched 2nd-harmonic generation by frequency-doubling of a high-frequency superimposed laser-diode *Opt. Lett.* **20** 273–275

[22] Watson M A, O'Connor M V, Lloyd P S, Shepherd D P, Hanna D C, Gawith C B E, Ming L, Smith P G R and Balachninaite O 2002 Extended operation of synchronously pumped optical parametric oscillators to longer idler wavelengths *Opt. Lett.* **27** 2106–2108

[23] Sturman B, Aguilar M, AgulloLopez F, Pruneri V and Kazansky P G 1997 Photorefractive nonlinearity of periodically poled ferroelectrics *J. Opt. Soc. Am. B* **14** 2641–2649

[24] Furukawa Y, Kitamura K, Alexandrovski A, Route R K, Fejer M M and Foulon G, 2001 Green-induced infrared absorption in MgO doped $LiNbO_3$ *Appl. Phys. Lett.* **78** 1970–1972

[25] Arbore M A, Marco O and Fejer M M 1997 Pulse compression during second-harmonic generation in aperiodic quasi-phase-matching gratings *Opt. Lett.* **22** 865–867

[26] Lefort L, Puech K, Butterworth S D, Svirko Y P and Hanna D C 1999 Generation of femtosecond pulses from order-of-magnitude pulse compression in a synchronously pumped optical parametric oscillator based on periodically poled lithium niobate *Opt. Lett.* **24** 28–30

[27] Laurell F and Karlsson H 1997 Electric field poling of flux grown $KTiOPO_4$ *Appl. Phys. Lett.* **71** 3474–3476

[28] Rosenman G, Skliar A, Eger D, Oron M and Katz M 1998 Low temperature periodic electrical poling of flux-grown $KTiOPO_4$ and isomorphic crystals *Appl. Phys. Lett.* **73** 3650–3652

[29] Jiang Q, Thomas P A, Hutton K B and Ward R C C 2002 Rb-doped potassium titanyl phosphate for periodic ferroelectric domain inversion *J. Appl. Phys.* **92** 2717–2723

[30] Chou M H, Parameswaran K R, Fejer M M and Brener I 1999 Multiple-channel wavelength conversion by use of engineered quasi phase-matching structures in $LiNbO_3$ waveguides *Opt. Lett.* **24** 1157–1159

[31] Amin J, Pruneri V, Webjorn J, Russell P S, Hanna D C and Wilkinson J S 1997 Blue light generation in a periodically poled Ti:$LiNbO_3$ channel waveguide *Opt. Commun.* **135** 41–44

[32] Hofmann D, Schreiber G, Haase C, Herrmann H, Grundkotter W, Ricken R and Sohler W 1999 Quasi-phase-matched difference-frequency generation in periodically poled Ti : $LiNbO_3$ channel waveguides *Opt. Lett.* **24** 896–898

[33] Risk W P, Gosnell T R and Nurmikko A V 2003 *Compact Blue–Green Laser* (Cambridge: Cambridge University Press)

[34] Bonfrate G, Pruneri V, Kazansky P G, Tapster P and Rarity J G 1999 Parametric fluorescence in periodically poled silica fibers *Appl. Phys. Lett.* **75** 2356–2358

[35] Wang W, Tavlykaev R and Ramasawamy R V 1997 Bandpass traveling-wave Mach–Zehnder modulator in $LiNbO_3$ with domain reversal *IEEE Photon. Technol. Lett.* **9** 610–612

[36] Yamada M, Saitoh M and Ooki H 1996 Electric-field induced cylindrical lens, switching and deflection devices composed of the inverted domains in $LiNbO_3$ crystals *Appl. Phys. Lett.* **69** 3659–3661

[37] Li J, Cheng H C, Kawas M J, Lambeth D N, Schlesinger T E and Stancil D D 1996 Electrooptic wafer beam deflector in LiTaO3 *IEEE Photon. Technol. Lett.* **8** 1486–1488

C1.1
Optical transmission

Michel Joindot and Michel Digonnet

C1.1.1 Introduction

The enormous potential of optical waves for high-rate transmission of information was recognized as early as the 1960s. Because of their very high frequency, it was predicted that lightwaves could be ultimately modulated at extremely large bit rates, well in excess of 100 Gbit s^{-1} and orders of magnitudes faster than possible with standard microwave-based communication systems. The promise of optical waves for high-speed communication became a reality starting in the late 1980s and culminated with the telecommunication boom of the late 1990s, during which time a worldwide communication network involving many tens of millions of miles of fibre was deployed in many countries and across many oceans. In fact, much of the material covered in this handbook was generated to a large extent as a result of the extensive optoelectronics research that was carried out in support of this burgeoning industry. The purpose of this chapter is to provide a brief overview of the basic architectures and properties of the most widely used type of optical transmission line, which exploit the enormous bandwidth of optical fibre by a general technique called wavelength-division multiplexing (WDM). After a brief history of optical network development, this chapter examines the various physical mechanisms that limit the performance of WDM systems, in particular, their output power (which affects the output signal-to-noise ratio (SNR)), capacity (bit rate times number of channels), optical reach (maximum distance between electronic regeneration) and cost. The emphasis is placed on the main performance-limiting effects, namely fibre optical nonlinearities, fibre chromatic and group velocity dispersions, optical amplifier noise and noise accumulation, and receiver noise. Means of reducing these effects, including fibre design, dispersion management, modulation schemes, and error-correcting codes, are also reviewed briefly. The text is abundantly illustrated with examples of both laboratory and commercial optical communication systems to give the reader a flavour of the kinds of system performance that are available. This chapter is not meant to be exhaustive, but to serve as a broad introduction and to supply background material for the following two chapters (Optical network architecture and Optical switching and multiplexed architectures), which dwell more deeply into details of system architectures. We also refer the reader to the abundant literature for a more in-depth description of these and many other aspects of optical communication systems (see, for example [1, 17, 32, 34]).

C1.1.2 History of the introduction of optics in backbone networks

Enabling the implementation of the optical communication concept required the development of a large number of key technologies. From the 1960s through the 1980s, many academic and industrial laboratories around the world carried out extensive research towards this goal. The three most difficult R&D tasks were the development of reliable laser sources and photo-detectors to generate and detect the optical signals, of suitable optical fibres to carry the signals, and of the components needed to perform such basic functions as splitting, filtering, combining, polarizing and amplifying light signals along the fibre network. Early silica-based fibres had a large core and consequently carried a large number of transverse modes, all of which travel at a different velocity, leading to unavoidable spreading of the short optical bits that carry the information and thus to unacceptably low bit rates over long distances. Perhaps, the most crucial technological breakthrough was the development of single-mode fibres, which first appeared in the mid-1970s and completely eliminated this problem. Over the following decade, progress in both material quality and manufacturing processes led to a dramatic reduction in the propagation loss of these fibres, from tens of dB per kilometre in early prototypes to the amazingly low typical current loss of 0.18 dB km^{-1} around 1.5 μm used in submarine systems (or an attenuation of only 50% through a slab of glass about 17 km thick!). The typical attenuation of fibres used in long-distance terrestrial networks today is around 0.22 dB km^{-1} at 1550 nm.

Fibre components were developed in the 1980s, including such fundamental devices as fibre couplers, fibre polarizers and polarization controllers, fibre wavelength-division multiplexers [5, 48], and rare-earth-doped fibre sources and amplifiers [17, 18, 41]. The descendants of these and several other components now form the building blocks of modern optical networks. Interestingly, the original basic research on almost all of these components was actually done not with communication systems in mind, but for fibre sensor applications, often under military sponsorship, in particular for development of the fibre optic gyroscope [5]. Parallel work on optoelectronic devices produced other cornerstone active devices, including high quantum efficiency, low-noise photo-detectors, efficient and low-noise semiconductor laser diodes in the near infrared, in particular distributed-feedback (DFB) lasers, as well as semiconductor amplifiers, although these were eclipsed in the late 1980s by rare-earth-doped fibre amplifiers. The development of high-power laser diodes, began in the 1980s to pump high-power solid-state lasers, in particular for military applications, sped up substantially in the late 1980s in response for the growing demand for compact pump sources around 980 and 1480 nm for then-emerging erbium-doped fibre amplifiers. Another key element in the development of optical communication networks was the advent of a new information management concept called SONET (Synchronous Optical NETworks) [31], especially matched to optical signals but also usable for other transmission technologies.

Up until the mid-1980s, long-distance communication network systems were based mostly on coaxial cable and radio frequency technologies. Although the maximum capacity of a single coaxial cable could be as high as 560 Mbit s^{-1}, most installed systems operated at a bit rate of 140 Mbit s^{-1}, while radio links could support typically eight 140 Mbit s^{-1} radio channels. Intercontinental traffic was shared between satellite links and analogue coaxial undersea systems; digital undersea coaxial systems never existed. The switch to optical networks was motivated in part by the need for a much greater capacity, in part by the need for improved security and reliability of radio-based and cable-based systems. These systems were commonly affected by two different types of failures, namely signal fading and cable breaks due to civil engineering work, respectively. The first optical transmission systems were introduced in communication networks in the mid-1980s. Early prototypes were classical digital systems with a capacity that started at 34 Mbit s^{-1} and rapidly grew to 140 Mbit s^{-1}, i.e. comparable to established technologies. Optical communication immediately outperformed the coaxial technology in terms of regeneration span, which was tens of kilometres compared to less than 2 km for high-capacity coaxial-cable systems. However, there was no significant advantage compared with radio links, in terms of either capacity or regeneration span length, the latter being typically around 50 km. One could thus envision future long-distance networks based on a combination of secure radio links and optical fibres. Soon after optical devices became reliable enough for operation in a submerged environment, optical fibre links rapidly replaced coaxial-cable systems. The very first optical systems used multimode fibres and operated around 800 nm. This spectral window was changed to 1300 nm for the second generation of systems, when lasers around this wavelength first became available. In Japan, where optical communication links were installed early on, prior to the development of the 1550-nm systems, many systems operate in this window. Most of the current systems for backbone networks, especially in Europe and the United States, operate in the spectral region known as the C-band (1530–1565 nm). This has become the preferred window of operation because the attenuation of silica-based single-mode fibre is minimum around 1550 nm. The first transatlantic optical cable, TAT-8, was deployed in 1988. Containing two fibre pairs and a large number of repeaters, it spans a distance of about 6600 km under the Atlantic Ocean between Europe and the USA and carries 280 Mbits of information per second. In 1993, optical transmission systems carrying 2.5 Gbit s^{-1} (16 × 155 Mbit s^{-1}) over a single fibre with a typical regeneration span of 100 km began to be added to the growing worldwide optic–optic network. In terms of both capacity and transmission quality, radio-based systems could no longer compete, and optics became the unique and dominating technology in backbone networks.

The single most important component that made high-speed communication possible over great distances ($\geqslant$100 km) without electronic regeneration is the optical amplifier. Although the loss of a communication optic around 1.5 μm is extremely small, after a few tens of kilometres, typically 50–100 km, the signal power has been so strongly attenuated that further propagation would cause the SNR of the signal at the receiver to degrade significantly, and thus the transmission quality, represented by the bit error rate, to be seriously compromised. The SNR can be improved by increasing the input signal power, but the latter can only be increased so much before the onset of devastating non-linear effects in the optic, in particular stimulated Raman scattering (SRS), stimulated Brillouin scattering (SBS) and four-wave mixing. Moreover, the gain in distance would be limited: a transmission over 200 km instead of 100 km of current optic would require the input power to be increased by roughly 20 dB!

This distance limitation was initially solved by placing optoelectronics repeaters along the optical line. Each repeater detects the incoming data stream, amplifies it electronically and modulates the current of a new laser diode with the detected modulation. The modulated diode's output signal is then launched into the next segment in the optic link. This approach works well, but its cost is high and its bit rate is limited, on both counts by the repeaters' high-speed electronics. A much cheaper alternative, which requires high-speed electronics only at the two ends of the transmission line, is optical amplification. Each electronic repeater is now replaced by an in-line optical amplifier, which amplifies the low-power signals that have travelled through a long optic span before their SNR gets too low, and then re-injects them into the next segment in the optic link. The advantage of this all-optical solution is clearly that the optical signal is never detected and turned into an electronic signal, until it reaches the end of the long-haul optical line, which can be thousands of kilometres long. Because the noise figure of optical amplifiers is low, typically 3–5 dB, the SNR can still be quite good even after the signals have travelled through dozens of amplifiers.

Starting as early as the 1960s, much research was devoted to several types of in-line optical amplifiers, first with semiconductor waveguide amplifiers [51], then with rare-earth-doped optic amplifiers [18], and more recently Raman optic amplifiers [28]. Semiconductor amplifiers turned out to have the highest wall-plug efficiency. However, at bit rates under about 1 Gbit s^{-1}, in WDM systems they induce cross-talk between signal channels. Although solutions have been recently proposed, semiconductor amplifiers have not yet entered the market in any significant way, partly because of the resounding success of the erbium-doped optic amplifier (EDFA). First reported in 1987 [42], this device provides a high small-signal gain around 1.5 μm (up to $\sim$50 dB) with a high saturation power and with an extremely high efficiency—the record is 11 dB of small-signal single-pass gain per milliwatt of pump power [52]. EDFAs used in telecommunication systems operate in saturation and have a lower gain, but it is still typically as high as 20–30 dB. The EDFAs can be pumped with a laser diode, at either 980 or 1480 nm, and they are thus very compact. Another key property is their wide gain spectrum, which stretches from $\sim$1475 to $\sim$1610 nm, or a total bandwidth of 135 nm ($\sim$16.4 THz!). For technical reasons, a single EDFA does not generally supply gain over this entire range, but rather over one of three smaller bands, called the S-band (for 'short', ~1480–1520 nm), the C-band (for 'conventional', ~1530–1565 nm) and the L-band (for 'long', ~1565–1610 nm). Amplification in the S-band can also be accomplished with a thulium-doped fibre amplifier (TDFA) [49]. Gain has been obtained over the S- and C-band by combining an EDFA and a TDFA [50].

Perhaps more importantly, EDFAs induce negligible channel cross-talk at modulation frequencies above about 1 MHz. These unique features make it nearly ideally suited for optical communication systems around 1.5 μm. Since the mid-1990s, it has been the amplifier of choice in the overwhelming majority of deployed systems, thus eliminating the electrical regeneration bottleneck.

The very large gain bandwidth of EDFAs and other optical amplifiers also provided the opportunity of amplifying a large number of modulated optical carriers at different wavelengths distributed over the amplifier bandwidth. This concept of wavelength division multiplexing had of course already been applied in radio links. One significant advantage of WDM optical systems is that the same amplifier amplifies many optical channels, in contrast with classical regenerated systems, which require one repeater per channel. Optical amplifiers thus reduce the installation cost of networks, in two major ways. First, the WDM technique results in an increase in capacity without laying new fibres, which reduces optic cost. Second, the cost of amplification is shared by a large number of channels, and because the use of a single optical amplifier is cheaper than implementing one regenerator per channel, the transmission cost is reduced proportionally. This critical economic advantage provided the final impetus needed to displace regenerated systems and launch the deployment of the worldwide optical WDM backbone networks that took place at the end of the 1990s.

C1.1.3 General structure of optical transmission systems

C1.1.3.1 Modulation and detection: RZ and NRZ codes

While radio systems use a wide variety of modulation formats in order to improve spectrum utilization, in optical systems data have been so far transmitted using binary intensity modulation. A logic 1 (resp. 0) is associated to the presence (resp. absence) of an optical pulse. Two types of line codes are mainly encountered: non-return to zero (NRZ), where the impulse duration is equal to the symbol duration (defined as the inverse of the data rate), and return to zero (RZ), where the impulse duration is significantly smaller than the symbol duration. This property explains why the name 'return to zero' is used; if the impulse duration equals roughly one half of the symbol time, the modulation format is designed as RZ 50%. So at a bit rate of 10 Gbit s^{-1}, NRZ uses impulses with a width of approximately 100 ps, while RZ 50% or RZ 25% will use 50 and 25 ps wide pulses, respectively. RZ has, for a given mean signal power, a higher signal peak power. This property can be used to exploit non-linear propagation effects, which under certain conditions can improve system performance. Details will be provided further on.

Research is actively being conducted to investigate new modulation schemes for future high-bit-rate systems. For instance, duobinary encoding, a well-known modulation scheme in radio systems, has been proposed because of its higher resistance to chromatic dispersion. Carrier-suppressed RZ (CSRZ), an RZ modulation format with an additional binary phase modulation, is also extensively studied, as well as recently differential phase shift keying (DPSK); both provide a higher resistance to non-linear effects. However, only NRZ and RZ are used in installed systems today.

Detection of the optical signals at the end of a transmission line is performed with a photo-detector, which is typically a PIN or an avalanche diode. Photons are converted in the semiconductor in electron–holes pairs and collected in an electrical circuit. The generated current is then amplified and sent to a decision circuit, where the data stream is detected by comparing the signal to a decision threshold, as in any digital system. Detection errors can occur in particular because of the presence of noise on the signal and in the detector. The error probability is a measurement of the transmission quality. In practice, the error probability is estimated by the bit error rate (BER), defined as the ratio of error bits over the total number of transmitted bits.

Several sources of noise are typically present in the detection process of an optical wave. Shot noise, the most fundamental one, arises from the discontinuous nature of light. Thermal noise is generated in the electrical amplifiers that follow the photo-detector. In PIN receivers, thermal noise is typically 15–20 dB larger than the quantum limit, and if the optical signal is low thermal noise dominates shot noise. In the case of amplified systems under normal operating conditions, the amplified spontaneous emission noise of the in-line amplifiers is largely dominant compared to the receiver noise, which can thus be neglected.

C1.1.3.2 Basic architecture of amplified WDM communication links

A typical amplified WDM optical link is illustrated in figure C1.1.1. The emitter consists of *N* lasers of different wavelengths, each one representing a communication channel. The lasers are typically DFB semiconductor lasers with a frequency stabilized by a number of means, including temperature control and often Bragg gratings. Each laser is amplitude modulated by the data to be transmitted. This modulation is performed with an external modulator, such as an amplitude modulator based on lithium niobate waveguide technology. Direct modulation of the laser current would be simpler and less costly, but it introduces chirping of the laser frequency, which is unacceptable at high modulation frequencies over long distances [1, 29]. The fibre-pigtailed laser outputs are combined onto the optical fibre bus using a wavelength division multiplexer, then generally amplified by a booster fibre amplifier.

The multiplexer can be based on concatenated WDM couplers (for low number of channels) or arrayed waveguide grating (AWG) multiplexers. This is the technology of choice for high channel counts, in particular in the so-called dense wavelength-division multiplexed (DWDM) systems: although this term has no precise definition, it applies usually to systems with a channel spacing less than 200 GHz. In-line optical amplifiers are distributed along the fibre bus to periodically amplify the power in the signals, depleted by lossy propagation along the fibre. Ideally, each amplifier provides just enough gain in each channel to compensate for the loss in that channel, i.e. such that each channel experiences a net gain of unity. Because the gain of an optical amplifier and to a smaller extent the loss

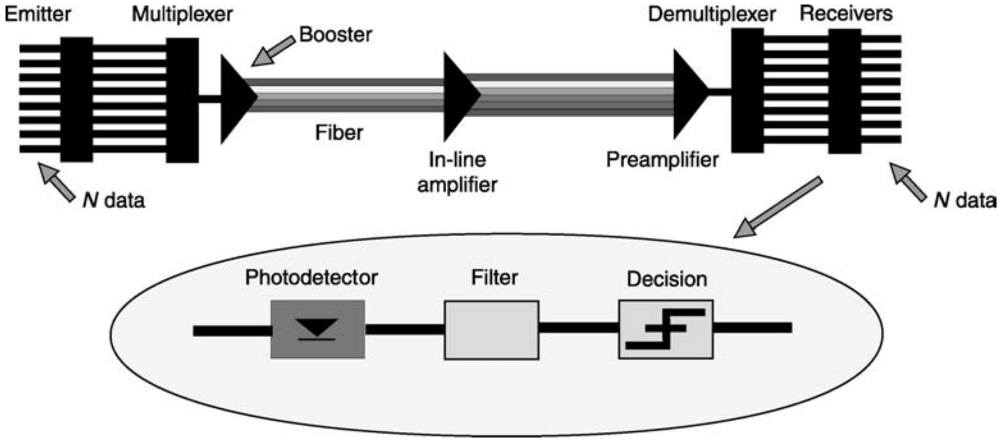


Figure C1.1.1. Structure of an amplified WDM optical system.

along the fibre are wavelength dependent, the net gain is different for different channels. If the difference in net gain between extreme channels is too large, after a few amplifier/bus spans the power in the strongest channel will grow excessively, thus robbing the gain for other channels and making their SNR at the receiver input unacceptably low. This major problem is typically resolved by flattening or otherwise shaping the amplifier spectral gain profile, or equivalently equalizing the power in the channels, using one of several possible techniques, either passive (for example long-period fibre gratings) [58] or dynamic (e.g. with variable optical attenuators). While the first WDM systems that appeared in the mid-1980s used basic amplifiers without control system, the new very long reach systems include complex gain flattening devices that compensate for the accumulation of gain tilt. The gain of an in-line amplifier ranges approximately from 15 to 30 dB per channel. The distance between amplifiers is typically 30–100 km, depending on fibre loss, number of channels and other system parameters. In deployed undersea systems, the amplifiers are equally spaced, whereas in terrestrial networks the amplifier location depends on geographical constraints, for instance, building location and amplifiers tend to be unevenly spaced. At the output of the transmission line, a wavelength division demultiplexer separates the *N* optical channels, which are then sent individually to a receiver, then electronically processed.

C1.1.3.3 Basic architectures of repeaterless systems

A repeaterless communication system aims to accomplish a very long optical reach without in-line amplifiers. A common application is connecting two terrestrial points on each side of a straight or narrow arm of sea, in which case it is generally not worth incurring the cost of undersea amplifiers. Some deployed repeaterless systems are extremely long, as much as hundreds of kilometres, and consequently they exhibit a high span attenuation, up to 50 or even 60 dB. The problem is then to ensure that the output power at the receiver is high enough, in spite of the high span loss, to achieve the required SNR at the end of the line.

This goal has been achieved with a number of architectures. A common solution involves using a preamplifier, i.e. an amplifier placed before the receiver to increase the detected power and reduce the receiver noise figure, as is also often done in classical WDM systems. Another one is to use a high-power amplifier, i.e. an amplifier placed between the emitter and the transmission fibre to boost the signal power launched into the fibre. Although such a booster amplifier is also present in some of the amplified WDM systems described before, the typical feature in repeaterless systems is the high power level, which can reach up to 30 dBm. In both cases, the amplifier can be either an EDFA or a Raman amplifier, or a combination of both. This solution, as we will see further on, is limited by non-linear effects in the fibre, although they can be somewhat mitigated with proper dispersion management.

A third solution specific to repeaterless systems is to place an amplifier fibre in the transmission fibre itself, and to pump it remotely with pump power launched into the transmission fibre from either end. The amplifier fibre can then be a length of Er-doped fibre; the entire transmission fibre can be lightly doped with erbium (the so-called distributed fibre amplifier); or the transmission fibre can be used as a Raman amplifier. The drawback of this general approach is that it requires a substantially higher pump power than a traditional EDFA, and it is therefore more costly. The reason is that the pump must propagate through a long length of transmission fibre before reaching the amplifier fibre, and because the transmission fibre is much more lossy at typical pump wavelengths than in the signal band, some of the pump power is lost. A fourth general solution, which is not specific to repeaterless systems, is to use powerful error-correcting schemes [12].

C1.1.3.4 Optical reach and amplification span

Two important features of a WDM communication system are its total capacity, usually expressed as $N \times D$, where N is the number of optical channels and D the bit rate per channel, and the optical reach, which is the maximum distance over which the signal can be transmitted without regeneration. Even in amplified systems with a nominally unity net gain transmission, due to the accumulation of noise from the optical amplifiers and signal distortions, after a long enough transmission distance the bit error becomes unacceptably high, and the optical signals need to be regenerated. In practical deployed WDM systems in 2001, this limitation typically occurs after about seven amplifier spans with a loss of roughly 25 dB per span.

Another important parameter is the amplification span, i.e. the distance between adjacent amplifiers. The performance of an optical WDM system cannot be expressed only in terms of optical reach; the number of spans must also be introduced. As an example, the optical reach of commercially available terrestrial systems in 2002 was around 800 km, compared to 6500 km in transatlantic systems. A key difference between them is the amplification span, as will be explained in the next section. In the following, WDM systems with a bit rate per channel of 2.5, 10 and 40 Gbit s^{-1} will be designated as *WDM* 2.5*G*, *WDM* 10*G* and *WDM* 40*G*, respectively.

C1.1.4 Limitations of optical transmission systems

C1.1.4.1 Noise sources and bit error rate

Amplifier noise

Amplification cannot be performed without adding noise to the amplified signals. In optical amplifiers, this noise originates from amplified spontaneous emission (ASE) [17], which is made of spontaneous emission photons emitted by the active ions (Er^{3+} in the case of EDFAs) via radiative relaxation subsequently amplified as they travel through the gain medium. The spectral power density of the ASE signal per polarization mode is given by:

$$\gamma_{\mathrm{ASE}} = n_{\mathrm{sp}} h\nu (G - 1) \tag{C1.1.1}$$

where G is the amplifier gain, h Planck's constant and ν the signal optical frequency. n_{sp} is a dimensionless parameter larger or equal to unity called the spontaneous noise factor. It depends on the amplifier's degree of inversion, and it approaches unity (lowest possible noise) for full inversion of the active ion population. The ASE is a broadband noise generated at all frequencies where the amplifier supplies gain, and its bandwidth is nominally the same as that of the amplifier gain. The ASE power coming out of the amplifier, concomitantly with the amplified signals, is obtained by the integration of γ_{ASE} over the frequency bandwidth of the gain. As an example, in a particular C-band EDFA amplifying ten signals equally spaced between 1531 and 1558 nm and with a power of 1 μW each, and with a peak gain of 33 dB at 1531 nm, the total power in the amplified signals is 5.5 mW, whereas the total ASE output power is 0.75 mW, i.e. more than 10% of the signals' power.

The photo-detectors used in receivers are the so-called quadratic detectors, i.e. they respond to the square of the optical field. Detection of an optical signal S corrupted by additive noise N (ASE noise in the case of amplified systems) in a photo-detector thus gives a signal proportional to $|S + N|^2$. Expansion of this signal gives rise to the signal $|S|^2$ (the useful signal) plus two noise terms. The first term ($2SN$) is the beat noise between the signal and the ASE frequency component at the signal frequency; it is called the signal–ASE beat noise. The second term (N^2) is the beat noise between each frequency component of the ASE with itself (the ASE–ASE beat noise). The signal–ASE beat noise varies from channel to channel, but the ASE–ASE beat noise is the same for all channels. A third and fourth noise terms are of course the shot noise of the amplified signal and the shot noise of the ASE, and to this must be added a fifth term, namely the receiver noise discussed earlier. In high-gain amplifiers with low input signals, which are applicable to most in-line amplifiers in communication links, the dominant amplifier noise term is the signal–ASE beat noise. In the amplifier example given at the end of the previous paragraph, the SNR degradation (also known as the noise figure) is $\sim 3.4\,\mathrm{dB}$ for all ten signals, and it is due almost entirely to signal–ASE beat noise. This noise term is typically large compared to the receiver noise, which can usually be neglected. Note that the NF is defined as the SNR degradation of a *shot-noise-limited* input signal. The SNR degradation at the output of an amplifier is therefore equal to the noise figure only when the input signal is shot-noise limited. In a chain of amplifier, this is true for the first amplifier that the signal traverses. However, after travelling through several amplifiers, the signal is no longer shot-noise limited but dominated by signal–ASE beat noise, and the SNR degradation is smaller than the noise figure. Refer to 'Accumulation of noise' section for further detail on noise accumulation in amplifier chains.

Photo-receiver thermal noise

As mentioned earlier, the photo-receiver thermal noise is generally fairly large compared to shot noise. However, it can become negligible when the signals are amplified with a preamplifier placed before the detector. To justify this statement, consider a receiver consisting of an optical preamplifier of gain G followed by a photo-detector. The optical signal and ASE noise powers at the receiver input are proportional to G and $G - 1$, respectively (in practice, G is very large and $G - 1 \approx G$). Because the thermal noise does not depend on G, and because it is typically 15–20 dB worse than the quantum limit, it is clear that if the preamplifier gain is large enough, say 20 dB, the thermal noise is negligible compared to the signal–ASE beat noise. This is exactly the same phenomenon as in electronics, where the high-gain first stage of a receiver masks the noise of the following stages. This property illustrates another advantage brought by optical amplifiers: optical preamplifiers allow to get away from the relatively poor noise figure of electronic circuits, and thus to achieve much better performance.

Relationship between bit error rate and noise

How does the error rate at the receiver depend on the noise level, or more exactly on the optical signal-to-noise ratio (OSNR), of the detected signal? To answer this question, we must make some assumptions regarding both the signal and the noise. First, because the signal–ASE beat noise depends on the signal power, it also depends on the state of the signal, i.e. on the transmitted data. If we assume an ideal on–off keying (OOK) modulation, signal–ASE beat noise is present only when the signal is on, whereas the ASE–ASE beat noise is present even in the absence of signal. Because the data can be assumed to be equally often on and off, the mean signal power is equal to half the peak power.

Second, to obtain an analytical expression for the bit error probability requires another assumption, common in communication theory, which is that the noise has a Gaussian statistics. This is true for signal–ASE beat noise, as a result of the linear processing of Gaussian processes, but it is not true for ASE–ASE beat noise. However, under normal operating conditions of amplified systems (i.e. with a sufficiently high OSNR), the influence of ASE–ASE beat noise remains relatively small. After a large number of optical amplifiers, however, the ASE–ASE beat noise component, which depends on the total ASE noise, can become significant. An effective way to reduce this noise component is then to place before the receiver an optical filter that cuts down the ASE power *between* the optical signals. This can be accomplished with a comb filter or with the demultiplexer that separates the channels. Such a filter reduces the ASE–ASE beat noise, but of course it does not attenuate either the signals or the ASE at the signals' frequencies, so it does not affect the signal–ASE beat noise. In the following, we assume that such a filter, with a rectangular transmission spectrum of optical bandwidth B_a, is placed before the receiver.

Third, since the noise variance is not the same conditionally to the transmitted data, the best decision threshold is not just at equal distance between the two signal levels associated with the two possible data values at the sampling time, but rather some other optimum threshold value that depends on signal power. Assuming that this optimum threshold value is used, the bit error probability (or bit error rate, BER) can be expressed as [30]:

$$P_{\text{exact}} = \frac{1}{2}\operatorname{erfc}\left(\frac{\sqrt{2}R}{\sqrt{m}+\sqrt{m+4R}}\right) \tag{C1.1.2}$$

where erfc is the complementary error function and the SNR R is the ratio of the mean signal power to the ASE power within the electrical bandwidth B, i.e. $\gamma_{\text{ASE}}B$. The electrical bandwidth B is the bandwidth of the electronic post-detection circuits. The parameter m is the normalized optical filter bandwidth, $m = B_a/B$. Equation (C1.1.2) can be easily derived by computing the variances of the signal–ASE beat noise and ASE–ASE beat noise contributions. For the computation of the first term, the average power of the signal is used. An ideal rectangular optical filter is assumed, as well as a rectangular electrical filter.

The SNR is usually measured not within the signal bandwidth, but over a much larger bandwidth B_0, corresponding generally to 0.1 nm in wavelength (or 12.5 GHz near 1550 nm). Calling this parameter the optical SNR R_0, the bit error probability can be rewritten as:

$$P_{\text{exact}} = \frac{1}{2}\operatorname{erfc}\left(\frac{Q}{\sqrt{2}}\right) \tag{C1.1.3}$$

where Q is the quality factor:

$$Q = \frac{2R_0}{\sqrt{\mu}+\sqrt{\mu+4\beta R_0}}. \tag{C1.1.4}$$

β is the ratio B/B_0 of the electrical to measurement bandwidths, $R_0 = \beta R$ and $\mu = m\beta^2$. An error probability of 10^{-9} (resp. 10^{-15}) requires for example $Q = 6$ (resp. 8). Neglecting the ASE–ASE beat noise contribution ($\mu \rightarrow 0$), the BER can be simply expressed as:

$$P_{\text{exact}} = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{R_0}{2\beta}}\right). \tag{C1.1.5}$$

From equation (C1.1.4), the value of R_0 needed to achieve a given quality factor Q_0 is given by:

$$R_0 = \beta Q_0^2\left(1+\frac{\sqrt{m}}{Q_0}\right). \tag{C1.1.6}$$

As an example, consider a 10 Gbit s^{-1} system with an optical bandwidth B_a of 50 GHz, an electrical bandwidth B of $\sim$ 7 GHz (as a rule of thumb, the electrical bandwidth is taken to be 70% of the bit rate), and a filter bandwidth B_0 of $\sim$ 12.4 GHz (0.1 nm). Then $\beta = 0.56$ and $m = 7$, and a BER of 10^{-15} ($Q_0 = 8$) requires an OSNR R_0 of $\sim$ 17 dB.

Based on the various degradation mechanisms that induce power penalty along the transmission system, equipment vendors specify a minimum SNR required by a given system. Typically, for *WDM* 10*G* systems, the minimum OSNR is in the range of 21–22 dB. It must be noted that the required OSNR increases with increasing electrical bandwidth, which is proportional to the bit rate. For instance, by going from 2.5 to 10 Gbit s^{-1} the OSNR needs to be increased by about 6 dB (see equation (C1.1.6)). This is a very important constraint, for system designers as well as operators.

Accumulation of noise

The accumulation of noise generated by successive amplifiers along an optical line degrades the OSNR. This accumulated noise limits the number of successive amplifiers that can be used, and thus the optical reach. Assuming a link with N equally spaced amplifiers of mean output power per channel P_0, an inversion parameter n_{sp}, and a gain G compensating exactly for the attenuation of the fibre span between them, the total noise power P_{n} (including both polarization modes) within a bandwidth B_0 at the last amplifier output is given by [29]:

$$P_{\mathrm{n}} = 2Nn_{\mathrm{sp}}(G-1)h\nu B_0 = n_{\mathrm{sp}}\frac{L}{Z_{\mathrm{a}}}(\exp(\alpha Z_{\mathrm{a}})-1)h\nu B_0 = 2n_{\mathrm{sp}}\alpha L\frac{G-1}{\ln G}h\nu B_0 \qquad \text{(C1.1.7)}$$

where L is the total length of the link, Z_{a} the length of the amplification span and α the attenuation factor of the fibre. The SNR R is:

$$R(N) = \frac{P_0}{2Nn_{\mathrm{sp}}(G-1)h\nu B_0} \qquad \text{(C1.1.8)}$$

or, in dB:

$$\mathrm{OSNR}\,(dB) = P_0\,(\mathrm{dBm}) - P_{\mathrm{n}}\,(\mathrm{dBm}). \qquad \text{(C1.1.9)}$$

These relationships show that for a given launched power and a given amplification span, the maximum transmission distance, represented by the maximum number of spans, is limited by the minimum required OSNR at the receiver input. For a given distance L, the OSNR increases when G decreases, i.e. when the amplification span becomes shorter. As an example, equation (C1.1.7) shows that for a link with a fixed length L and a fibre attenuation of 0.2 dB km^{-1}, an OSNR improvement of 7 dB is achieved when the span length is reduced from $Z_{\mathrm{a}} = 100$ km (a span loss of 20 dB) to $Z_{\mathrm{a}} = 50$ km (a span loss of 10 dB). The link with the 50-km span length does require twice as many amplifiers, but the gain they each need to supply is reduced by 10 dB. So is their output noise power, and as a result the OSNR is improved. This illustrates how important a parameter the fibre attenuation is. For a given amplification span and a given number of spans, any reduction in this attenuation will result in a better OSNR simply because the amplifier gain G will be smaller. Equivalently, a lower fibre loss allows increasing the optical reach. For example, a fibre loss reduction as small as 0.02 dB km^{-1}, from 0.23 to 0.21 dB km^{-1}, in 100-km spans, will allow to reduce the gain by 2 dB and thus to improve the OSNR by as much as 2 dB. If with the 0.23 dB km^{-1} fibre the required OSNR was reached after eight spans, with the 0.21 dB km^{-1} fibre it will be possible to have 12 spans ($10\log(12/8) = 1.7$ dB), the same noise power being then produced by a larger number of less noisy amplifiers. This dependence of the OSNR on the span loss explains why the amplification span is significantly shorter in undersea lightwave systems compared to terrestrial ones, because transmission distances are much longer.

Figure C1.1.2 shows the output signal power per channel (calculated from equation (C1.1.8)) needed to reach an OSNR of 20 dB as a function of the number of spans for different span losses. The inversion parameter of the optical amplifiers is taken to be $n_{\mathrm{sp}} = 1.6$, and the output ASE is assumed to be filtered with a 0.1-nm narrowband filter ($B_0 = 12.5$ GHz). These curves also allow comparing the power required for achieving a given optical reach. For instance, for a link length $L = 1000$ km and a fibre attenuation $\alpha = 0.25$ dB km^{-1}, if there are $N = 20$ spans, each one of them will have a loss $\alpha L/N = 12.5$ dB, and figure C1.1.2 shows that the required power per channel will be -7.5 dBm.

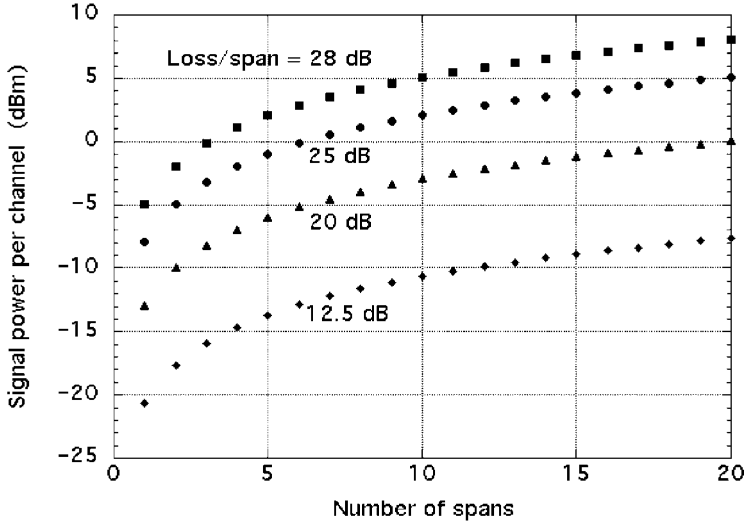


Figure C1.1.2. Output signal power per channel required to achieve an OSNR of 20 dB as a function of the number of spans for different span losses.

When the number of spans is divided by two ($N = 10$), the span loss increases to 25 dB and the required signal power jumps nearly tenfold, to 2 dBm.

C1.1.4.2 Signal distortions induced by propagation

In addition to SNR degradation due to optical amplifiers, transmitted optical signals also suffer distortions induced by propagation along the fibre. These effects become more and more important as the bit rate increases.

Chromatic dispersion

Within a narrow bandwidth around the carrier angular frequency ω_0, a fibre of length L can be viewed as an all-pass linear filter, with an attenuation nearly independent of wavelength over the small signal bandwidth prevailing at the bit rates under consideration. Expanding the phase up to the second order in frequency ω about ω_0 allows to write the transfer function $H(\omega)$ of this filter as:

$$H(\omega) = A\exp\left[\mathrm{i}\left(\Phi(\omega_0) + \beta_1 L(\omega - \omega_0) + \frac{1}{2}\beta_2 L(\omega - \omega_0)^2\right)\right]. \tag{C1.1.10}$$

Higher order terms in the expansion must be considered when for instance β_2 equals zero. The first and second terms of the exponent represent a constant phase shift and the delay of the impulse (also called group delay), respectively. The third term, proportional to the second derivative of the signal mode index with respect to the wavelength, originates physically from the dependence of the mode group velocity on wavelength. It is often referred to as group velocity dispersion (GVD), or chromatic dispersion. In an optical waveguide such as a fibre, GVD is approximately the sum of the material dispersion and the fibre dispersion. It is mathematically represented by β_2, usually expressed in $\mathrm{ps}^2\,\mathrm{km}^{-1}$, or by the so-called chromatic dispersion D, which is a more familiar parameter to system designers, expressed in $\mathrm{ps\,nm}^{-1}\,\mathrm{km}^{-1}$. It is the group delay variation over a 1-nm bandwidth after propagation along a 1-km length of fibre. In a standard communication fibre, D is around $15\,\mathrm{ps\,nm^{-1}\,km^{-1}}$. D (expressed in $\mathrm{ps\,nm^{-1}\,km^{-1}}$) is related to β_2 (in $\mathrm{ps^2\,km^{-1}}$) and λ (in nm) by:

$$D(\lambda) = -6\pi 10^5 \frac{\beta_2(\lambda)}{\lambda^2}. \tag{C1.1.11}$$

Chromatic dispersion results in broadening of the signal as it propagates through the fibre. When the signal pulse amplitude is Gaussian, the pulse width evolution along the fibre can be computed analytically and pulse broadening can be expressed with simple expressions [1]. If a Gaussian pulse with a complex impulse envelope $u(t, 0)$ of temporal width θ_0

$$u(t, 0) = U_0 \exp\left(-\frac{t^2}{2\theta_0^2}\right) \tag{C1.1.12}$$

is launched into the fibre at $z = 0$, the impulse at distance L is given by [1]:

$$u(t, L) = U \exp\left(-\frac{T^2}{2\theta^2(x)} + \mathrm{i}\Psi(x, t)\right) \tag{C1.1.13}$$

where U is the amplitude taking into account the fibre attenuation, $T = \mathrm{i}t - \beta_1 L$ the time in local coordinates associated to the signal and $\Psi(x, T)$ a phase term. The parameter $\theta(x)$ is the temporal pulse width at distance L, given by:

$$\theta(x) = \theta_0 \sqrt{1 + x^2} \tag{C1.1.14}$$

where $x = L/L_\mathrm{D}$ is the propagation distance normalized to the characteristic dispersion length $L_\mathrm{D} = \theta_0^2/|\beta_2|$.

As expected from physical arguments, in the presence of chromatic dispersion the pulse width expands along the fibre in much the same way as a spatial beam expands in space due to diffraction (see equation (C1.1.14)). Here the dispersion length L_D plays the same role as the Rayleigh range does in the diffraction of Gaussian beams. For a fibre of length L and dispersion coefficient β_2, there is an optimum value of the incident pulse width θ_0 that minimizes the pulse width at the fibre output. This optimum pulse width, obtained by setting the derivative of equation (C1.1.14) with respect to θ_0 equal to zero, is given by:

$$\theta_{0,\mathrm{opt}} = \sqrt{|\beta_2| L} \tag{C1.1.15}$$

and the output pulse width is:

$$\theta(L) = \sqrt{2}\theta_{0,\mathrm{opt}} = \sqrt{2|\beta_2| L}. \tag{C1.1.16}$$

Stated differently, equation (C1.1.15) shows that for a given input pulse width θ_0 the optimum fibre length that minimizes the output temporal pulse width is $L = \theta_0^2/|\beta_2| = L_\mathrm{D}$, i.e. one dispersion length.

This analysis shows that the larger the chromatic dispersion is, the narrower the initial pulse needs to be, and the larger the pulse width will be at the output of the fibre. It implies that if the input pulse width is not properly selected, i.e. if it is either too narrow or too wide, chromatic dispersion will cause successive pulses to overlap, which creates what is known as intersymbol interference (ISI). This deleterious effect alters the decision process and thus increases the bit error rate. It must be noted that a Gaussian pulse extends indefinitely in time and there is theoretically always a finite amount of ISI; the maximum distance is thus set by a 'tolerable' level of ISI. This distance is exactly L_D if we define this acceptable ISI is reached when the initial width of the pulse has been multiplied by $\sqrt{2}$ (which is somewhat arbitrary). As an example, assuming Gaussian pulses with θ_0 equal to half the symbol duration, for a standard communication fibre ($D = 17\,\mathrm{ps\,nm^{-1}\,km^{-1}}$, i.e. $\beta_2 = 20\,\mathrm{ps^2\,km^{-1}}$) this distance is equal to 2000 km at 2.5 Gbit $\mathrm{s^{-1}}$, but only 125 km at 10 Gbit $\mathrm{s^{-1}}$.

This result demonstrates that propagation at bit rates of 10 Gbit $\mathrm{s^{-1}}$ or greater in a standard fibre is not possible over distances longer than a few tens of km without significant ISI. This problem is circumvented in practice by introducing along the transmission line components with a negative dispersion coefficient to compensate for chromatic dispersion, in much the same way as an optical lens is used to refocus a free-space beam after it has expanded as a result of diffraction. This method has been demonstrated in the laboratory with a number of optical filters, especially fibre Bragg gratings for dynamic compensation [19], or more simply and commonly with a length of dispersion-compensating fibre (DCF) designed to exhibit a strong negative dispersion coefficient D [17], in the range of $-90\,\mathrm{ps\,nm^{-1}\,km^{-1}}$ for standard fibre. This last solution is the only one used in commercial systems today, and fibre suppliers try to develop the best compensation fibre matched to the fibre they sell. DCFs are typically more lossy than standard communication fibres, with attenuation coefficients around $0.5\,\mathrm{dB\,km^{-1}}$. To make up for this additional loss, the DCF is typically inserted near an amplifier. In order to reduce the impact of the DCF loss on the amplifier noise figure, in WDM systems the DCF is usually placed in the middle of a two-stage amplifier.

Chromatic dispersion depends on wavelength. This dependence is characterized by the dispersion slope, expressed in $\mathrm{ps\,nm^{-2}\,km^{-1}}$). In order to completely compensate for the dispersion at any wavelength, the fibre and the associated DCF must exhibit the same D/S ratio, where S is the slope of the dispersion D. The existence of a perfectly slope-compensating DCF depends strongly on the type of fibre. For example, a DCF very well matched to standard single-mode fibre is available, but this is not true for all fibres. If the slope is not matched, some channels will exhibit a finite residual chromatic dispersion outside the ‘acceptance window’, i.e. the interval within which the dispersion must lie to ensure a correct transmission quality. This window is typically $1000\,\mathrm{ps\,nm^{-1}}$ wide for a *WDM* 10*G* system. In general, unless carefully designed a dispersion-compensation filter does not cancel dispersion perfectly for all channels. So even after correcting dispersion to first order, in long-haul WDM systems the residual dispersion can still limit the transmission length and/or the number of channels. To illustrate the magnitude of this effect, consider a link of length L carrying N channels spaced by $\Delta\lambda$ (i.e. $N\Delta\lambda$ is the total multiplexed width), with a dispersion slope after compensation S. The cumulated dispersion at the output of the link is then $SLN\Delta\lambda$. The receiver can be designed to tolerate a certain amount of residual cumulated dispersion within some spectral window, for instance typically $1000\,\mathrm{ps\,nm^{-1}}$ for a 10 Gbit $\mathrm{s^{-1}}$ system, as stated earlier. For a typical dispersion slope $S = 0.08\,\mathrm{ps\,nm^{-2}\,km^{-1}}$, $N = 64$ and $\Delta\lambda = 0.8\,\mathrm{nm}$ (or a multiplexed width of 51.2 nm), the maximum possible fibre length for which the cumulated dispersion reaches $1000\,\mathrm{ps\,nm^{-1}}$ is 244 km. This effect can of course be avoided by reducing the length or the number of channels, which impacts system performance. Better solutions include designing broadband dispersion-compensation filters with a dispersion curve matched to that of the fibre link. This is a key issue for WDM systems, which has received a lot of attention from system designers.

Nonlinear effects

The maximum power that can be transmitted through an optical fibre, and thus the SNR of the signal at the fibre output, are ultimately limited by a number of optical nonlinearities present in the optical fibre. These non-linear effects are the Kerr effect (dependence of the fibre refractive index on the signal intensity), stimulated Brillouin scattering (conversion of signal power into a frequency-shifted backward wave), stimulated Raman scattering (conversion of signal power into a forward frequency-shifted wave) and four-wave mixing (optical mixing of a signal with itself or other signals and concomitant generation of spurious frequencies). The magnitude of these effects generally increases with increasing signal intensity, which can be relatively high in a single-mode fibre, even at low power, because of the fibre's large optical confinement. For example, in a typical single-mode fibre at 1.55 μm with an effective mode area of 80 μm^2, a 20-dBm signal has an intensity of $\sim 1.2\,kW\,mm^{-2}$! Because this intensity is sustained over very long lengths, and because the conversion efficiency of non-linear effects generally increases with length, even the comparatively weak non-linear effects present in silica-based fibres can have a substantial impact on system performance, even at low power. This section provides background on the magnitude of these non-linear effects, describes their impact on system performance, and mentions typical means of reducing them.

Self-phase modulation (SPM)

When a signal propagates through an optical fibre, through the Kerr effect it causes a change Δn in the refractive index of the fibre material. In turn, this modification of the medium property reacts on the signal by changing its velocity and thus its phase. This non-linear effect is known as SPM. For a signal of power P, the index perturbation Δn is expressed as:

$$\Delta n = n_2 \frac{P}{A_{\text{eff}}} \tag{C1.1.17}$$

where n_2 is the Kerr non-linear constant of the fibre ($n_2 \approx 3.2 \times 10^{-20}\,m^2\,W^{-1}$ for silica) [35] and A_{eff} the signal mode effective area. The resulting change in the mode propagation constant β is:

$$\Delta\beta = \frac{\omega \Delta n}{c} = \frac{2\pi n_2}{\lambda} I \tag{C1.1.18}$$

where $I = P/A_{\text{eff}}$ is the signal intensity. In the case of a modulated signal, because the Kerr effect has an extremely fast response time ($\ll 1$ ps) each portion of the signal pulse modulates its own phase independently of other portions of the pulse. If $I_0(t)$ is the instantaneous intensity, or equivalently the intensity profile, of the signal launched into the fibre, and if α is the fibre loss at the signal wavelength, then the signal intensity at a point z along the fibre is $I(t,z) = I_0(t)\exp(-\alpha z)$. The amount of SPM experienced by the signal pulse after a propagation length L is simply [1]:

$$\Phi(t,L) = \int_0^L \Delta\beta\,dz = \frac{2\pi n_2}{\lambda}\int_0^L I_0 \exp(-\alpha z)\,dz = \frac{2\pi n_2 I_0}{\lambda} L_{\text{eff}} \tag{C1.1.19}$$

where $L_{\text{eff}} = (1 - e^{-\alpha})/\alpha$ is the effective fibre length.

In principle, because photo-detectors are quadratic and thus phase insensitive, SPM should not cause any detrimental effect. This is true in AM systems provided the fibre is free of dispersion. However, in practice, the presence of chromatic dispersion converts SPM into amplitude fluctuations [1]. When β_2 is positive, SPM combines to chromatic dispersion to produce pulse broadening, just like chromatic dispersion alone does. When β_2 is negative, SPM combines to chromatic dispersion to produce pulse narrowing, i.e. they have opposite effects. In this case, SPM can be used to compensate for chromatic dispersion and thus improve the system performance. There is in fact a particular regime in which linear and non-linear effects compensate mutually exactly at any moment in time. This particular solution of the non-linear Schrödinger equation, valid only in a lossless fibre, is called an optical soliton. Provided that it has the proper shape and intensity, a soliton propagates without any temporal deformation. This phenomenon was extensively studied in the 1990s, and it continues to be an active research topic, because fibre-optic solitons are very promising for ultra-long distance transmission at high bit rates [26, 43, 44]. A soliton-based transmission encrypts the information in extremely short pulses that neither spread nor compress as they propagate along the fibre because the soliton has just the right peak power for the Kerr non-linear phase shift to exactly compensate for chromatic dispersion. Soliton-based communication links are, however, not compatible with WDM-based links because in order to have a relatively low peak power, a soliton needs a low-dispersion fibre, which is not well suited for WDM (see 'Four-wave mixing (FWM)' section). The WDM solution has obviously won so far, even for long-haul transmission. But the concept of soliton-based communication systems remains an interesting and promising approach that continues to stimulate a lot of research and development.

In parallel to these various schemes used to combat SPM, the most effective first-order solution to reduce SPM is to use a transmission fibre with a large mode effective area A_{eff}. This is of course also applicable to other undesirable fibre non-linear effects, in particular cross-phase modulation, four-wave mixing and stimulated scattering processes. The reason is that the magnitude of all of these processes increases as the reciprocal of A_{eff}, so a fibre with a higher A_{eff} can tolerate a higher signal power. Large mode effective areas are typically accomplished by designing fibres with a larger core and a concomitantly lower numerical aperture to ensure that the fibre carries a single mode. Communication-grade fibres have mode effective areas in the range of $50-100\,\mu\mathrm{m}^2$, for example $80\,\mu\mathrm{m}^2$ for the so-called standard fibre. Substantially higher values are typically precluded for transmission fibres because they require such low numerical apertures that the fibre becomes overly susceptible to bending loss.

Cross-phase modulation (XPM)

XPM has the same physical origin as SPM, namely the Kerr effect, except that the phase modulation is not induced by a signal on itself, but by one or more different signals propagation through the fibre. A different signal means any signal with a different wavelength, a different polarization, and/or a different propagation direction. In a WDM system, the phase of a signal of wavelength λ_i is therefore modulated by itself (wavelength λ_i, SPM) and by all the other channels (wavelengths $\lambda_{j\neq i}$, XPM).

The XPM affecting a particular channel i of a WDM system depends on the power (and therefore on the data) and wavelength of all other channels $j \neq i$. As in the case of SPM, XPM is converted into amplitude fluctuations through chromatic dispersion. However, the main detrimental effect of XPM is time jitter, due to the fact that the other signals also change the group delay of channel i. The position of the impulses is thus changed randomly around an average position, and sampling before decision does not occur always at the same instant within the pulse, which cause a BER penalty. If we consider the case of one interfering channel, interaction occurs when two pulses overlap. Because they propagate at different speeds (the group velocities at the wavelengths of the two channels), the interaction begins when the fastest impulse starts to overlap with the slowest pulse and ends when it has completely passed it. This phenomenon is called a collision. After one collision between symmetrical pulses, there is theoretically no memory on the perturbed pulse. The problem occurs in the case of an incomplete collision, for instance, when it begins just before an amplifier and then the powers change during the collision. In this case, the affected pulse keeps the memory through a shifted temporal position. A key parameter is to characterize this effect is the difference of group velocity between the two channels, which is equal to $D\Delta\lambda$ where D is the dispersion and $\Delta\lambda$ the channel spacing. If this parameter is high, the effect will be smaller, because collisions will be very rapid. Increasing the channel spacing will then reduce the interaction because the difference between group velocities is larger. The influence of chromatic dispersion is more complex. A higher dispersion reduces channel interaction and thus phase modulation, but as discussed earlier it also increases conversion into amplitude fluctuations. Further details can be found in section C1.1.5.

Four-wave mixing (FWM)

FWM is another non-linear process that results directly from the Kerr effect. Channels of a WDM system beat together in the receiver, giving rise to intermodulated side-bands at frequencies that are sums and differences of the channel's frequencies. Each of these side-bands is modulated with the information encrypted on the channels that gave rise to it. When a side-band frequency happens to fall on or close to one of the channel's frequencies, this channel becomes modulated with unwanted information from other channels. This intermodulation has the same undesirable side effects as similar effects well known in radio systems.

As an illustration, consider a communication system utilizing channels that are equally spaced in frequency, which is usual in deployed systems, i.e. the channel frequencies are $f_0 + m\Delta f$, where m is an integer. The third-order beating between channels 0, 1 and 2 at respective frequencies f_0, $f_1 = f_0 + \Delta f$ and $f_2 = f_0 + 2\Delta f$ produces side-band signals at frequencies $pf_0 + qf_1 + rf_2$, where $|p| + |q| + |r| = 3$. In particular, an intermodulated side-band is generated at frequency $f_0 + \Delta f$ by interaction of three channels together ($p = 1$, $q = -1$, $r = -1$), but also by interaction of channels 0 and 1 only ($p = 1$, $q = 2$). This side-band has the same frequency as channel 1, and thus adds to channel 1 data modulation from channels 0 and 2. The same argument applied to other channels clearly shows that if the interaction is strong enough, every channel becomes contaminated with information from all other channels.

The magnitude of FWM effects can be characterized by the power in the intermodulation side-band P_{intermod}. This power can be calculated analytically for pure unmodulated waves, in which case it is given by [15, 56]:

$$P_{\text{intermod}} = \eta_{\text{FWM}} d^2 \gamma^2 P^3 \exp(-2\alpha L) \tag{C1.1.20}$$

where $\gamma = 2\pi n_2/\lambda A_{\text{eff}}$ represents the strength of the Kerr nonlinearity in the fibre, P is the power per channel, assumed the same for all channels, d is a constant equal to 6 if all channels are distinct and 9 if there are not. The factor η_{FWM} is the FWM efficiency, defined as:

$$\eta_{\text{FWM}} = \frac{\alpha^2}{\alpha^2 + \Delta\beta_{\text{FWM}}^2}\left[1 + \frac{4\exp(-\alpha L)\sin^2(\Delta\beta_{\text{FWM}}L/2)}{(1 - \exp(-\alpha L))^2}\right] \tag{C1.1.21}$$

where $\Delta\beta_{\text{FWM}}$ is the phase mismatch between interacting waves, which depends on chromatic dispersion coefficient D, on its slope, and on the channel spacing Δf according to:

$$\Delta\beta_{\text{FWM}} = \frac{2\pi\lambda^2}{c}\Delta f^2\left(D + \Delta f\frac{\lambda^2}{c}\frac{\partial D}{\partial\lambda}\right). \tag{C1.1.22}$$

In the usual case where the total attenuation of the span is high enough ($\exp(-\alpha L) \ll 1$), the efficiency (equation (C1.1.21)) is well approximated by:

$$\eta_{\text{FWM}} = \frac{\alpha^2}{\alpha^2 + \Delta\beta_{\text{FWM}}^2}. \tag{C1.1.23}$$

FWM is a phase-matched process: for energy to flow effectively from one channel to another, the channels must remain in phase, i.e. the phase mismatch $\Delta\beta_{\text{FWM}}$ must be small. It means that the closer the channel frequencies are (small Δf), the more efficient FWM is, as indicated mathematically by equations (C1.1.21) and (C1.1.22). This explains why the intermodulation power increases with decreasing channel spacing. Chromatic dispersion plays a beneficial role by increasing the phase mismatch between channels and thus reducing the FWM efficiency, as shown by equation (C1.1.22).

The intermodulation power also increases with increasing channel power, and it does so rapidly (as the third power in P) because FWM is a non-linear process.

Figure C1.1.3 shows the effect of both dispersion and channel spacing on the interference-to-carrier ratio, i.e. the difference between the channel power and the intermodulation product power. This quantity is plotted versus channel spacing for four values of the dispersion typical for channels located near the zero-dispersion wavelength. This figure simulates a fibre link with a length $L = 100\,\mathrm{km}$, a fibre attenuation of $0.2\,\mathrm{dB\,km^{-1}}$, a dispersion slope of $0.08\,\mathrm{ps\,nm^{-2}\,km^{-1}}$, a non-linear coefficient $\gamma = 3\,\mathrm{W^{-1}\,km^{-1}}$, and a launched power per channel of 4 dBm. It is clear that a higher dispersion reduces FWM and thus allows a better utilization of the available bandwidth. For example, if a ratio of -60 dB is required, figure C1.1.3 shows that this can be accomplished with a 100-GHz channel spacing in a standard fibre ($D = 17\,\mathrm{ps\,nm^{-1}\,km^{-1}}$), but only 210 GHz or higher in a typical non-zero-dispersion-shifted fibre (family G.655) with a chromatic dispersion of 3 ps/(nm km).

In single-channel transmission, a low dispersion is beneficial because it reduces the amount of pulse spreading induced by (1) dispersion and (2) SPM combined with dispersion, and thus it reduces the amount of dispersion compensation needed to correct for these effects. In multi-channel transmission the situation is not as simple because dispersion now brings protection against inter-channel effects, XPM and FWM. But the situation depends strongly on the channel spacing: for *WDM* 10*G* systems with a typical channel spacing of 100 GHz or less, inter-channel effects are dominant compared to intra-channel effects (SPM). This is the reason why a dispersion-shifted fibre with zero dispersion around 1550 nm is much worse for WDM transmission than a standard G.652 fibre, and also why this fibre provides the smallest channel spacing at this bit rate (25 GHz). When higher bit rates are considered, the channel spacing cannot be reduced so much due to the spectral width of the modulated signals, and then intra-channel cannot be neglected compared to inter-channel effects.

Stimulated Brillouin scattering

Stimulated Brillouin scattering belongs to the family of parametric amplification processes. Through interaction between the optical signal and acoustic phonons, it causes power conversion from the signal into a counterpropagating signal shifted in frequency by the acoustic phonon frequency [1]. The power in the SBS signal grows as $\exp(g_B P_p - \alpha)z$, where g_B is the SBS gain, in $\mathrm{W^{-1}\,km^{-1}}$,

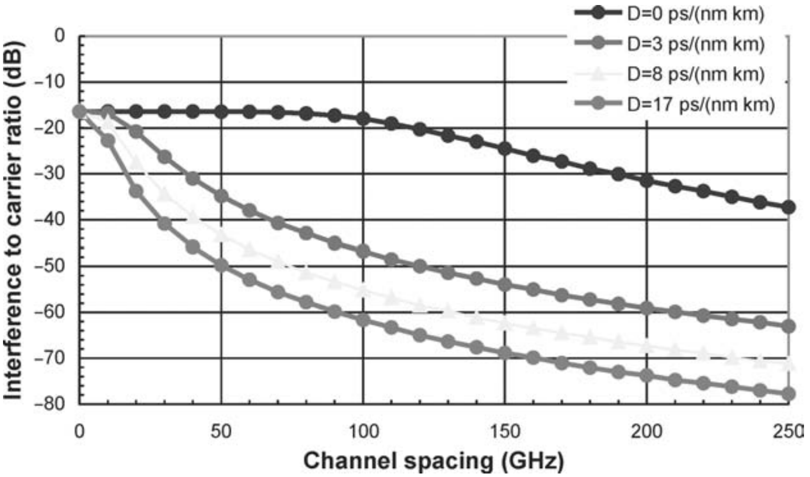


Figure C1.1.3. Dependence of the interference-to-carrier ratio due to four-wave mixing on the channel spacing.

which depends on the wavelength separation between the two signals, and α is the attenuation of the medium. SBS is a narrowband process. In a silica fibre, the Brillouin frequency shift at 1.55 μm is $\nu_B \approx 11$ GHz and the gain bandwidth is only $\Delta\nu_B \approx 100$ MHz. It is customary to characterize SBS by its power threshold, i.e. the power required to compensate for the medium attenuation and thus just begin to provide a positive gain. For an unmodulated signal with a linewidth smaller or equal to the SBS gain bandwidth, the SBS threshold in a typical 1.5-μm fibre (mode effective area of 50 μm^2) is around 21 mW km, i.e. 21 mW in a 1-km fibre and 2.1 mW in a 10-km fibre [1]. For powers larger than the threshold, a fraction or all of the signal is converted in the backward SBS signal. It is therefore essential to keep the power in each signal below the SBS threshold, and because this threshold is fairly low, SBS limits the power of a narrowband signal that can be transported over a given distance. This is particularly critical in repeaterless systems [25].

Several solutions have been demonstrated and are routinely applied to increase the Brillouin threshold and thus increase the power and or distance over which signal can be transported. When the signal amplitude is modulated at bit rates higher than ~100 MHz, as is the case in WDM systems, the signal bandwidth power exceeds the Brillouin linewidth and SBS is reduced. For this reason, SBS is generally not a concern in WDM systems. One caveat is that the carrier component of the modulated signal retains the original linewidth of the unmodulated signal, and it is still backscattered by SBS. Because the carrier component carries only half the signal average power, the SBS threshold is increased (by 3 dB) compared to an unmodulated narrowband signal, but in OOK schemes using high powers SBS acting on the carrier has been observed to induce signal distortion [33].

Because the SBS gain decreases with increasing carrier linewidth $\Delta\nu$ as $\Delta\nu_B/(\Delta\nu_B + \Delta\nu)$, another solution to further increase the SBS threshold is to use a larger carrier linewidth. This can be done with a directly modulated laser (direct modulation tends to chirp the laser frequency), by applying to the laser either a phase modulation [25, 36] or a small amount of frequency modulation (at a frequency much lower than the bit rate) [45]. Other techniques include using a duobinary modulation scheme to suppress the carrier component [37], concatenating fibres with different Brillouin shifts to reduce the interaction length [40], and placing isolators along the fibre to periodically suppress the backward SBS signal [54].

Stimulated Raman scattering

Although caused by a different physical mechanism (interaction with vibrational modes of the medium structure instead of acoustic phonons), stimulated Raman scattering can be modelled in a very similar manner, but its characteristics are quite different and so are its effects on transmission systems [1]. SRS is an optical process that causes power transfer between an optical pump and a co- or counter-propagating signal. Most solid media exhibit SRS, including silica-based fibres. *Spontaneous* Raman scattering occurs when a pump photon of frequency ω_p is scattered by a host phonon of frequency Ω, which results in the annihilation of the pump photon and the spontaneous emission of a signal photon at a frequency $\omega_s = \omega_p - \Omega$. This scattering process can also be stimulated when an incident signal photon of frequency ω_s interacts with a pump photon and a phonon, thus yielding the emission of a stimulated photon at frequency ω_s. This stimulated process thus provides what is known as Raman gain. The SRS gain spectrum is centred around a frequency downshifted from the pump frequency ω_p by the mean phonon frequency Ω of the material. The Raman gain spectrum and bandwidth are set by the finite-bandwidth phonon spectrum of the material. The Raman shift of silica is typically 13 THz (or ~100 nm at 1550 nm), which is much higher than for stimulated Brillouin scattering. Similarly, the gain full width at half maximum is larger, around 8 THz (70 nm for a pump around 1.55 μm). However, the Raman gain coefficient for a silica fibre is much weaker than the SBS gain, by a factor of about 500, so the Raman threshold is typically much higher, for example around 1.2 W in a 10-km length of 1.55-μm communication fibre with a 50-μm^2 effective mode area [1]. Although much weaker, SRS can still be deleterious in WDM systems because optical channels located at the highest gain frequencies act as pumps and can be depleted, while other channels can be amplified. In conventional systems using only the C-band (30 nm wide), SRS does not occur because the maximum separation between channels is much smaller than the Raman shift. But in systems using both the C- and L-bands, power transfer between channels of C- and L-bands can be induced by SRS and must be taken into account in system design.

It must be noted that SRS is also a useful mechanism: a pump signal injected in the fibre can transfer its power via SRS to one or more signals and thus provide amplification. This is the basic principle of fibre Raman amplifiers, which will be considered in section C1.1.6.4.

Polarization-mode dispersion (PMD)

A standard single-mode optical fibre does not actually carry a single mode but two modes with orthogonal, nearly linear polarizations. Because the index difference between the fibre core and the cladding is small, these two polarization modes have nearly degenerate propagation constants. However, these propagation constants are not exactly the same. In a communication link, the signal launched into the fibre is typically linearly polarized. The fibre exhibits random linear and circular birefringence, and as the signal propagates through it the signal polarization evolves through many states. Because the two orthogonal polarization modes travel at slightly different velocity, one lags behind the other, and because the signal is temporally modulated into short pulses, after long-enough propagation each pulse is split into two pulses. This produces two electrical pulses with amplitudes that depend on the polarization of the optical signal at the receiver, and separated by a random delay called differential group delay (DGD). For fibres with strong coupling between polarization modes, DGD follows a Maxwell distribution. The mean value of DGD is called the PMD [3, 16, 20, 22, 23].

This multipath effect causes ISI and thus degrades the bit error rate. Furthermore, random variations in the birefringence of the long fibre cause the DGD to be a random variable and thus the properties of the transmitted signals to be time dependent. Communication systems must then be characterized by their outage probability, or outage time, i.e. the probability that the BER exceeds the maximum tolerable value, above which transmission is no longer possible with the required quality. PMD is a linear effect, which, just like chromatic dispersion, acts on each channel individually but does not cause coupling between them.

WDM systems are usually designed to tolerate a PMD approximately equal to one-tenth of the symbol duration, or around 10–12 ps for a 10 Gbit s^{-1} bit rate. When this value is exceeded by a small amount, transmission can still be sustained with fewer channels, which allows increasing the SNR of the remaining channels and provides a better resistance against PMD. When PMD is too high (for instance 20 ps or more for a *WDM* 10*G* system), distortion can cause closure of the eye-diagram and increasing the power does not bring any improvement. Currently manufactured fibres allow 10 Gbit s^{-1} transmission over several thousands of kilometres, and PMD is not a problem at this bit rate. Recent advances in manufacturing processes have led to fibres with low enough PMD values for 40 Gbit s^{-1} transmission over more than 2000 km.

PMD compensation has been investigated in several laboratories, for example using feedback equalizers in either the optical or the electrical domain [46, 59]. The main application was the implementation of *WDM* 10*G* systems in existing fibre links that could originally not support this high bit rate because the fibre exhibited a high PMD. Although this method was successful, its economic viability has been questioned because it requires one equalizer per channel and its high cost cannot be shared. PMD compensation will certainly need to be implemented in the future in communication systems with higher bit rates over long distances, which have a reduced tolerance to PMD. For example, a *WDM* 40*G* system typically requires no more than 2 or 2.5 ps of PMD.

C1.1.5 Design of an optical WDM system

C1.1.5.1 Global performance of a system: BER and OSNR

As mentioned earlier, the performance of a WDM system is expressed in terms of its BER, which is obtained by measuring the number of error bits occurring over a given time interval. A minimum OSNR value is required in order to achieve the required transmission quality, i.e. the BER needs to be lower than a given threshold. Commercial equipment is typically specified in terms of OSNR: the maximum number of spans is specified for different losses per amplification span. For example, an OSNR of 22 dB will be guaranteed for seven spans of 25-dB loss (7 × 25 dB), or for ten spans of 23-dB loss (10 × 23 dB).

Another important feature is the system sensitivity to chromatic dispersion and dispersion-compensation strategy. The residual dispersion at the receiver input must remain within some interval. As an example, a $10\,\mathrm{Gbit\,s^{-1}}$ receiver will only accept a cumulated dispersion between -600 and $+800\,\mathrm{ps\,nm^{-1}}$. For a particular channel, it is always possible to bring the cumulated dispersion within this range with proper in-line compensation. However, due to the finite dispersion slope, the other channels will experience a different cumulated dispersion, and if the link is too long and/or the dispersion slope is too high, it will not be possible to meet this specification for all channels. This limitation could be lifted by adjusting the cumulative dispersion channel by channel, but this is not practical for economic reasons. As a result, chromatic dispersion generally imposes an upper limit on the bit rate and the optical reach.

C1.1.5.2 Critical parameters and trade-offs for terrestrial, undersea and repeaterless systems

As discussed earlier, the optical reach and amplification span are critical parameters in communication systems. For a given optical reach and a fixed launched power, a shorter amplification span improves the OSNR. Conversely, for a given required OSNR it increases the optical reach. However, a shorter amplification span also results in a more expensive system, more complex monitoring, and a higher operating cost. Moreover, in a terrestrial network the location of the amplification sites and the network topology in general are parameters that the operator does not want to change. The network infrastructure and fibres are long-term investments, and they are required to be compatible with several generations of systems. In particular, the attenuation per span is a constrained parameter. Its value is imposed by the characteristics of networks where systems have to be installed, and it is typically in the range of 20–25 dB. Technical improvement goals in terrestrial WDM systems therefore consist in increasing the capacity and the optical reach within the framework of this attenuation per span.

Undersea systems benefit from an additional degree of freedom. Unlike in terrestrial networks, the fibre and the system are laid together and cannot be separated, and therefore the cable does not need to be designed for successive generations of systems. The amplification span can then be selected without any location constraint. The preferred solution is to space the amplifiers equally because it simplifies manufacturing, dispersion compensation and cable maintenance. Because of the very long optical reach of undersea systems (for example 6500 km between Europe and North America and 9000 km across the Pacific Ocean), the amplification span must be substantially reduced, down to around 40 km.

In principle, the OSNR can always be improved by increasing the power launched into the fibre. However, the maximum *available* power is limited by the high cost of high-power amplifiers and also by safety rules. Quite independently, as discussed earlier the maximum *usable* power is limited by non-linear effects in the fibre (SPM, XPM and FWM) to a level that insures that these effects are maintained at or below some acceptable level. Two strategies are possible to take into account this power limitation. The first strategy is to reduce the power level when the number of spans is increased, because impairments caused by non-linear effects are cumulative along the transmission line. Systems using this type of approach are called linear, or also NRZ (because linear systems use NRZ pulses, i.e. pulses with a duration of roughly one symbol time). Another strategy is to exploit the beneficial effect of ‘soliton-like’ propagation regimes, where linear and non-linear effects cancel each other, which requires a precise compensation map. These systems are usually called non-linear, or RZ (because they used pulses of the RZ type, i.e. significantly shorter than the symbol time and a higher peak power than NRZ pulses).

In summary, the amplifier gain is determined by the span loss that needs to be compensated, the output power is limited by the cost and technology of amplifiers and by fibre non-linear effects, and the optical reach is then given by the minimum SNR that can be achieved.

C1.1.6 State of the art and future of the WDM technology

C1.1.6.1 State-of-the-art WDM system capacity and distance

The first WDM systems appeared in 1995 and could transmit four 2.5 Gbit s^{-1} channels. The number of channels was rapidly increased to 16, then 32, by reducing the frequency spacing between channels down to 100 GHz (0.8 nm). WDM systems were predicted to evolve towards even higher channel counts by reducing this spacing to 0.4 nm, but this change has taken place slowly because of the difficulty and higher cost of developing components, in particular multiplexers and demultiplexers, capable of handling signals so closely spaced in wavelength. Instead, in the next generation of systems the capacity was increased by increasing the bit rate to 10 Gbit s^{-1}. *WDM* 10*G* systems provide a higher capacity, but they are also subject to more severe propagation impairments. Chromatic dispersion, which can be neglected at 2.5 Gbit s^{-1} up to around 800 km, becomes a critical issue at higher bit rates, and dispersion compensation units must be incorporated in all amplification sites. In 2001, the state of the art for typical engineering data provided to system operators by suppliers was 80 × 10 Gbit s^{-1} channels with a 50-GHz spacing, or a total capacity of 800 Gbit s^{-1}, over a single fibre with an optical reach of around 700 km. Subsequent progress in filtering and laser technology has allowed the achievement of 25-GHz channel spacing, offering a capacity of 1.6 Tbit s^{-1} over the 30-nm-wide C-band of EDFAs. The capacity of commercially available fibre links has therefore been multiplied by a factor 160 in 7 years, and the symbolic barrier of 1 Tbit s^{-1} has already been exceeded. Furthermore, much higher capacities have been demonstrated in laboratories, for example up to 5 Tbit s^{-1} over 12 × 100 km [7] and even 10.2 Tbit s^{-1} over 100 km using polarization multiplexing [9].

Figure C1.1.4 depicts the state of the art of *WDM* technology at the time of this writing. Circles represent laboratory prototypes, and squares represent either commercially available or announced systems. The diagonal line represents the 1 Tbit s^{-1} capacity boundary; systems that fall in the hatched quadrant above it have a capacity greater than 1 Tbit s^{-1}. The highest commercial capacity is 3.2 Tbit s^{-1} (80 × 40 Gbit s^{-1}). Systems with a 1.6 Tbit s^{-1} capacity have been proposed in two configurations, namely 160 channels at 10 Gbit s^{-1} with a 25-GHz spacing, and 40 channels at 40 Gbit s^{-1} with a 100-GHz spacing [38]. Demonstrations have also been performed at 20 Gbit s^{-1} per channel, but no commercial system has yet been developed at this bit rate.

In spite of these great advances in experimental communication links, most of the links deployed in the world operate at much lower bit rates. For example, in France most links operate at 2.5 Gbit s^{-1}. The network is being upgraded to 10 Gbit s^{-1}, but it will take several years before the conversion is complete [47]. Operational systems in the USA are further along; most of the deployed links run at 10 Gbit s^{-1}. Very few commercial systems operate at 40 Gbit s^{-1}. Again the main difficulty is that these systems require more precise correction for chromatic dispersion and dispersion slope. It means using accurately tailored and stable fibre Bragg gratings so that dispersion does not take over. The system is then less lenient on imperfections in dispersion compensation circuits, and more difficult to develop and manage.

It is significant that all nodes in deployed terrestrial networks are opaque, i.e. the optical signals are detected, turned into electrical signals, amplified, switched (routed) electronically into the right direction

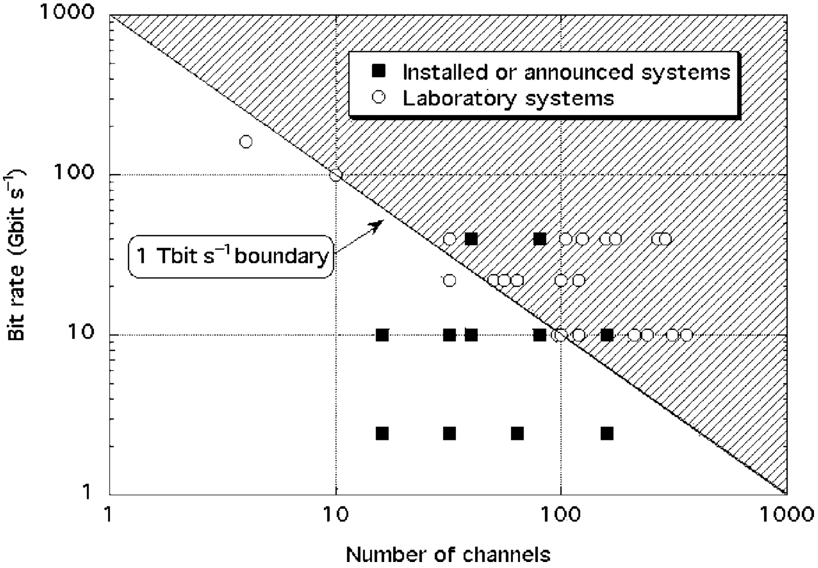


Figure C1.1.4. State of the art of WDM technology (see text for details).

and turned back into a light signal with a local laser oscillator. Such optoelectronic nodes constitute a large fraction of the cost of the network. The reason why this function is not all done all optically is the high cost, and to some extent loss, of optical components, in particular optical switches. Only a few companies have developed and are planning on deploying commercial transparent networks using all-optical nodes. The only market at the moment is secure military lines in the USA. Future WDM systems with extended optical reach (ULH and VLH, see section C1.1.6.3) are an efficient way to reduce the number of opto-electronic regeneration sites and thus reduce the network cost.

C1.1.6.2 Forward error-correcting codes

As explained earlier, the compatibility of network topology with future system generations is a very strong requirement for terrestrial networks. For example, upgrading an existing link from 2.5 to 10 Gbit s^{-1} requires an OSNR increase of 6 dB, which in turn should require a corresponding increase in power at the receiver. Instead of increasing the power, which again is not always possible for reasons covered earlier, a solution is the use of a forward error-correcting code [13], an approach widely used in satellite and radio systems. Redundancy is introduced in the input data (prior to transmission) and used by an electronic decoder to detect and eventually correct some of the errors affecting the data sequence through the detection process. A coding–decoding scheme is characterized by the proportion of errors it can detect and correct, which is related to the redundancy. An error-correcting code therefore allows the receiver to accept a smaller OSNR, which results in a degraded bit-error rate that is corrected by the decoder. The price to pay is an increase in transmitted data resulting from the required redundancy.

A very simple example is the repeat code: binary bits of information are repeated an odd number of times, and the decoder corrects for errors by comparing the like-bits of replica and retaining the bits that occur most often (a process known as majority decoding). This is not a very efficient code, because it requires a large amount of redundancy, but other, far more powerful codes exist, such as the well-known Reed–Solomon and BCH codes, that can reduce the bit error rate substantially at the cost of surprisingly low redundancy [13]. Most of the *WDM* 10*G* systems today utilize a Reed–Solomon code, while other more powerful schemes have been extensively studied for the next generations of systems [2, 27]. To illustrate the improvement brought by forward-error correction, a bit error rate of 10^{-5} at the input of a Reed–Solomon decoder results into a bit error rate of roughly 10^{-15} at its output. Assuming this last value is required by the user, it means that the transmission systems is only required to achieve a BER of 10^{-5} and can then operate at a lower OSNR. This approach allows in particular to gain the 6-dB difference in OSNR between *WDM* 2.5*G* and 10*G* systems and thus to operate a 10 Gbit s^{-1} system with the same OSNR as a 2.5 Gbit s^{-1} system, at the cost of a modest redundancy (or data rate increase) of ~7%. Forward error-correcting codes can also clearly be used in 2.5 Gbit s^{-1} systems to increase the optical reach or to operate with higher span losses. Forward error-correcting codes are likely to play a major role in the higher speed communication systems of the future.

C1.1.6.3 Ultra-long-haul technology: new problems arising

One key advantage of the WDM technology is that the high cost of amplifiers can be shared between all transmitted channels. Any reduction in the number of opto-electronic regeneration sites, or in other words any increase in the optical reach, is then very attractive. This is one major reason for the interest in ultra long haul (ULH) and very long haul (VLH) systems, which can operate today at 10 Gbit s^{-1} per channel over respective distances of ~ 1500 and ~ 3000 km. Raman amplifiers are a key element in this technology: they allow to improve the OSNR for a given distance or equivalently to increase the optical reach for a given OSNR. The compatibility of ULH and VLH systems with existing infrastructures is also a difficult challenge because the OSNR decreases as the number of spans increases, and thus the accumulation of propagation impairments is much more critical. This limitation is of course mitigated in practice with error-correcting codes, as well as Raman amplifiers.

C1.1.6.4 Raman amplification

Raman amplification relies on SRS, a non-linear optical process that causes power transfer between an optical pump and a signal (see 'Stimulated Raman Scattering' section). One of the greatest strengths of Raman amplification is that it can supply gain at any wavelength provided a suitable pump source is available. Since the Raman shift of silica is typically 13 THz, to obtain a gain peak at 1530 nm the pump wavelength must be ~1430 nm. With such a pump, gain will be available from about 1490–1546 nm. The Raman gain cross-section of silica is unfortunately relatively small, so the power requirement is much higher than for an EDFA, but Raman fibre amplifiers present several important benefits that somewhat mitigate this disadvantages, including the flexibility of pump wavelength selection, the availability of gain anywhere where pump is available, and the fact that the gain medium is the transmission fibre itself.

Raman fibre amplifiers can be configured in a number of ways, each with its own benefits and applications. Forward-pumped Raman amplifiers, in which the pump and the signals to be amplified travel in the same direction, induce cross-talk between channels, which is undesirable in WDM systems. This is the reason why the backward-pumped Raman fibre amplifier is often preferred. This configuration is illustrated in figure C1.1.5. An obvious advantage of the Raman amplifier is that it can be easily implemented in any existing fibre link by simply adding a pump source at the proper wavelength. In particular, it does not require the insertion of a doped fibre, which keeps the cost down. Also, in the event of pump failure the fibre amplifier is still transparent, as opposed to an EDFA where an unpumped Er-doped fibre is essentially opaque at the signal wavelengths. We refer the reader to chapter A1.6 for a more detailed description of Raman amplifiers.

To illustrate the system benefits of in-line Raman amplifiers, figure C1.1.6 shows the calculated evolution of signal power with distance from the emitter with and without Raman amplification. This

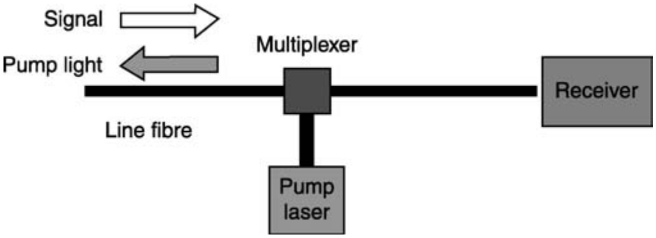


Figure C1.1.5. Schematic of a backward-pumped Raman fibre amplifier.

figure simulates a link with a fibre attenuation of 0.22 dB km^{-1}, a total length of 100 km, and a Raman gain of 10 dB around 1.53 μm. In the absence of amplification, the signal power in dBm decreases linearly with distance. When pump is injected at the far end of the fibre, SRS adds gain at that end of the fibre. The gain decreases away from the pump, as a result of pump photons being either scattered by the fibre or converted into signal photons. There is practically no more incremental Raman gain after the pump has travelled about 60 km, because by then most of it has been consumed. The system advantage of Raman amplification is to increase the signal power at the receiver and thus to improve the SNR of the detected signal. As with any amplifier, the greatest benefit to the SNR occurs when the signal is amplified before its power becomes too low. To understand this basic principle, consider two configurations. Configuration (a) is a link of length *L* with an amplifier of gain *G* placed at the end, and configuration (b) is the same link with the same amplifier of gain *G* placed a distance *d* before the receiver. In both configurations, the signal power at the receiver is the same, but in (b) the noise power is reduced by a factor equal to the loss *A* of the fibre length *d*. The in-line amplifier of configuration (b) is therefore equivalent to an amplifier located at the receiver with a noise figure reduced by *A*.

A backward-pumped Raman amplifier can be viewed as an in-line amplifier located before the receiver. The gain that it supplies allows (1) reducing the launched signal power and thus reducing non-linear effects for a given topology (span loss and number of spans); (2) increasing the span loss for a given number of spans while maintaining the same OSNR; (3) increasing the number of spans for a given span loss while maintaining again the same OSNR. Raman amplification is therefore a key technology to increase the optical reach and upgrade existing networks to higher bit rates.

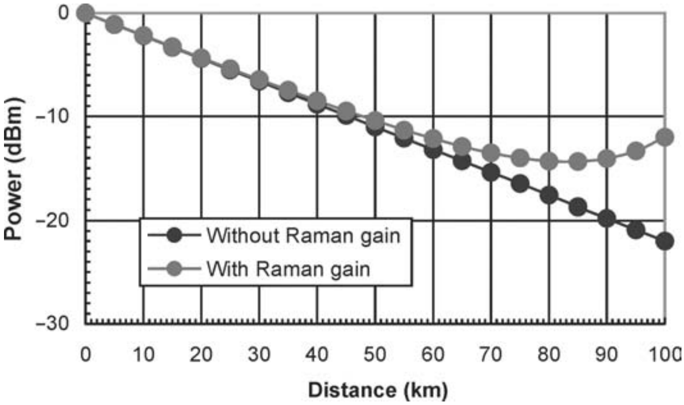


Figure C1.1.6. Typical variation of signal power versus distance with and without a fibre Raman amplifier.

C1.1.6.5 Diversification of fibres and ITU fibre standards

We have seen how propagation phenomena become more and more critical as the bit rate increases. These phenomena depend on several system parameters, such as the channel power and channel spacing, as well as fibre parameters, especially scattering loss, chromatic dispersion and effective mode area. It was therefore important, early on, to develop standard fibres with set ranges for all of their critical parameters. This task was accomplished by the International Telecommunications Union (ITU), which defined a number of other important standards and terminologies as well, such as the so-called ITU grid, i.e. the precise values of the discrete wavelengths used in optical communication systems around the world. One of the most important single-mode standard fibres is the G.652 fibre (also known as standard single-mode fibre, or SSMF), designed for early single-channel transmission in the 1550-nm band before the emergence of WDM and EDFAs.

Because the high dispersion of the G.652 fibre (17 $\mathrm{ps\,nm^{-1}\,km^{-1}}$) could be viewed as a drawback for single-channel transmission, the standard dispersion-shifted fibre (DSF), known as G.653, was designed to achieve zero dispersion at 1550 nm and thus drastically reduce signal distortion. As we saw earlier, a DSF is not well suited for WDM operation, because channels located near the zero-dispersion wavelength suffer severe non-linear effects, especially FWM. G.653 fibre can be used in WDM systems but at the price of significant constraints, such as increased and eventually irregular channel spacing. This results in a poorer spectral efficiency and thus a higher cost, which is the reason why it is generally not used in this application.

A third family of fibres, the non-zero-dispersion-shifted fibres (NZDSF), also known as G.655, became available somewhat later. Their main feature is a chromatic dispersion lower than that of the G.652 fibre. The initial idea was to choose a dispersion value high enough to keep deleterious non-linear effects at a low level while requiring less dispersion compensation. Identifying the right trade-off between dispersion and dispersion compensation was not an easy task. The chromatic dispersion of commercially available G.655 fibres has increased progressively, from 3 $\mathrm{ps\,nm^{-1}\,km^{-1}}$ in early fibres to 8 $\mathrm{ps\,nm^{-1}\,km^{-1}}$ for recent fibres. The dispersion of the NZDSF family of fibres therefore covers a wide range. Although G.652 is certainly the most widely used fibre in backbone networks, a considerable amount of G.655 fibres has been deployed, especially by new operators, in the second half of the 1990s. At a bit rate of 10 $\mathrm{Gbit\,s^{-1}}$ per channel, the advantage over G.652 fibres is not obvious; excellent performance has been reached with G.652 fibres because their higher dispersion allows very efficient protection against interchannel effects, and thus a closer channel spacing. The question of fibre choice is more critical and certainly still an open issue at higher bit rates (40 $\mathrm{Gbit\,s^{-1}}$ and above), although *WDM 40G* transmission has been successfully demonstrated over G.652 fibres in many laboratories.

C1.1.6.6 Towards the future: WDM 40G systems and beyond

There is a definite economic advantage in operating transmission systems at higher bit rates, in part because integration reduces cost, at least when a certain level of production is reached and technology is stabilized. For instance, a 10 $\mathrm{Gbit\,s^{-1}}$ emitter is less expensive than four 2.5 $\mathrm{Gbit\,s^{-1}}$ emitters, yet both provide the same capacity. The same is true for receivers. A drawback of moving towards higher bit rates is that the narrower impulses required are more sensitive to propagation phenomena, and also that the implementation of electronic circuits is more difficult. Another consideration is the channel spacing. After its amplitude has been modulated, the optical signal has a much wider spectrum than the original signal produced by the emitter. The linewidth of the signal from a typical emitter is of the order of a few MHz, whereas after modulation this linewidth becomes comparable to the modulation frequency, i.e. many GHz or tens of GHz. Maintaining a tolerable level of cross-talk then requires a higher channel spacing: for instance, a bit rate of 40 $\mathrm{Gbit\,s^{-1}}$ is certainly not compatible with a 25-GHz channel spacing.

However, with a 100-GHz channel spacing a 40 Gbit s^{-1} bit rate becomes possible. A sample of recent research in 40 Gbit s^{-1} systems is given in 'Selected recent results' section.

In response to the growing demand for higher speed communications, extensive research efforts have been expanded in industrial laboratories towards the development of *WDM* 40*G* systems, which are considered to be the next generation for optical data transmission. The main difficulties, as stressed earlier, are that this type of system has a much tighter tolerance to most design parameters, including chromatic dispersion, FWM, and polarization mode dispersion, and more accurate forms of compensation will be required. In addition, advanced research is already being conducted on even higher bit rates per channel, 80 and even 160 Gbit s^{-1} [53, 61]. The objective of this research is primarily knowledge acquisition. At this point it is difficult to predict whether these systems will ever be viable and in what time frame, and whether they will be cost-effective compared to existing generations.

Increasing the number of channels by increasing the amplification bandwidth

For a given channel spacing, or when the minimum channel spacing imposed by the bit rate is reached, it is still possible to increase the capacity by increasing the number of channels. The channels used in practice fall in one of three bands, which are defined by the amplification bands of EDFAs, as opposed to other requirements. The C-band (1530–1565 nm) is the region where the EDFA is most efficient, i.e. where it provides the most gain per unit pump power. This is the band that is used first, and in many installed networks it is the only band that is used. The L-band (1565–1610 nm) is used when the C-band is full and additional channels need to be added, even though L-band EDFAs are less efficient and provide a lower gain than C-band EDFAs. Generally, an EDFA cannot be optimized to provide gain efficiently in both bands, so an in-line amplification site for both C- and L-bands typically consists of two separate EDFAs, one operating in the C-band and the other one in the L-band. The two amplifiers are often placed in a parallel configuration, as illustrated in figure C1.1.7. A demultiplexer separates the incoming WDM signals into C-band signals, which are sent to the C-band EDFA, and L-band signals, which are sent to the L-band EDFA. A multiplexer placed at the output of the two EDFAs recombines the amplified signals onto the same fibre. Serial configurations are also possible. These configurations are not specific to the L- and C-bands: they can be used to multiplex any bands, in principle in any number. It should be pointed out that the economical advantage of using both the C- and L-bands in the same system is not obvious: amplification must be performed in two separate amplifiers, so that this technique allows only sharing the communication fibre. The benefits of using the L-band are also mitigated by other considerations. First, the fibre loss is higher in the L-band, and the C-band/L-band

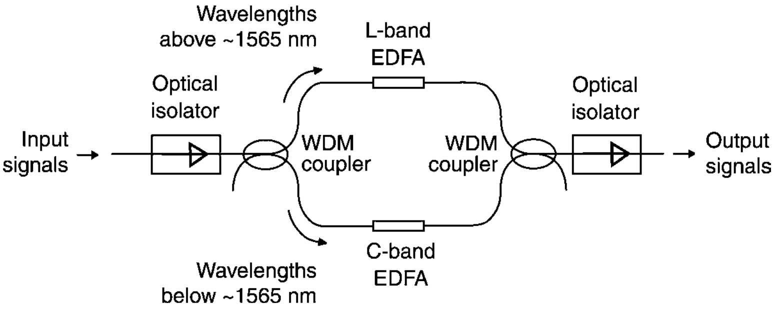


Figure C1.1.7. Diagram of a parallel arrangement of L-band and C-band EDFA.

multiplexers introduce losses as well, in both bands, which degrade the power budget and the OSNR. The trade-off between these extraneous losses and OSNR degradation on one end, and the cost of deploying additional fibre on the other, needs to be carefully examined. Second, L-band amplification is more costly than C-band amplification. The reason is that it requires a different set of components than C-band amplifiers, and these components are more expensive because they do not benefit from the same economy of scale. As a result, very few systems in the world use the L-band, except in Japan where new technologies tend to be deployed earlier than elsewhere and where existing fibres happen to have the right amount of GVD for WDM communication in the L-band. Use of the L-band is often a good solution for G.653 fibres: chromatic dispersion is higher in the L-band than in the C-band, and it provides some protection against deleterious non-linear effects such as FWM.

Very recently, EDFAs have also been designed to provide gain in the S-band (~1480–1530 nm) [4]. Although the gain and the gain efficiency in the S-band are even lower than in the L-band, they are both respectable, and the S-band is a possible future spectral window to extend the usable bandwidth of silica fibres. The S-band is in fact already used in metropolitan communication systems without amplification. However, amplification in the S-band is costly, even more than in the L-band. No deployed system uses amplification in the S-band, except perhaps isolated experimental systems. Because of the higher costs of L- and S-bands, communication companies generally find it economically preferable to use up the capacity of the C-band in a given link before starting to utilize other bands.

Increasing the number of channels with a closer channel spacing

The above considerations show that filling the C-band is one of the most efficient solutions for increasing the bandwidth. The channel spacing has been normalized by ITU to 100 GHz, but sub-grids with a spacing of 50 GHz and even 25 GHz can be used. $N \times 10\,\mathrm{Gbit\,s^{-1}}$ systems available today have commonly 100 and 50-GHz channelling, and 25 GHz has been proposed [38]. As is well known in the radio domain, a convenient parameter to describe the fraction of the available bandwidth that is actually utilized is the spectral efficiency. It is defined as the total transmitted bit rate (number of channels times channel bit rate) divided by the occupied bandwidth. For example, a $16 \times 10\,\mathrm{Gbit\,s^{-1}}$ link with a channel spacing of 100 GHz has a total capacity of $160\,\mathrm{Gbit\,s^{-1}}$ and occupies a bandwidth of 16×100 GHz, so the spectral efficiency is $0.1\,\mathrm{bit\,s^{-1}\,Hz^{-1}}$. The highest value obtained so far in commercially available optical systems based on OOK modulation is around $0.4\,\mathrm{bit\,s^{-1}\,Hz^{-1}}$ for $10\,\mathrm{Gbit\,s^{-1}}$ with a 25-GHz channel spacing in the C-band, and $0.8\,\mathrm{bit\,s^{-1}\,Hz^{-1}}$ has been demonstrated in the laboratory. These values remain modest compared to the spectral efficiencies of 4 or $5\,\mathrm{bit\,s^{-1}\,Hz^{-1}}$ currently achieved in radio systems. Such high values are made possible by multilevel modulation schemes. Similar schemes also exist in optics [6], but they have not yet been implemented in commercial systems.

Figure C1.1.8 depicts the evolution of the spectral efficiency of commercially available WDM systems since 1994. Higher values have been achieved in the laboratory, up to $1.28\,\mathrm{bit\,s^{-1}\,Hz^{-1}}$ [9]. Reducing the channel spacing has required several technological improvements over the years, including increasing the stability of demultiplexing filters and of the channel wavelengths (to this end, lasers are now equipped with wavelength lockers). The ultimate achievable channel spacing depends strongly on the fibre. As mentioned earlier, G.652 fibre is a very good candidate because its high chromatic dispersion reduces the threshold of FWM.

Modulation schemes

Optical systems use currently an OOK intensity modulation format. The advantage of this modulation scheme is that it is rather simple, and that it does not require excessive signal power. However, its main

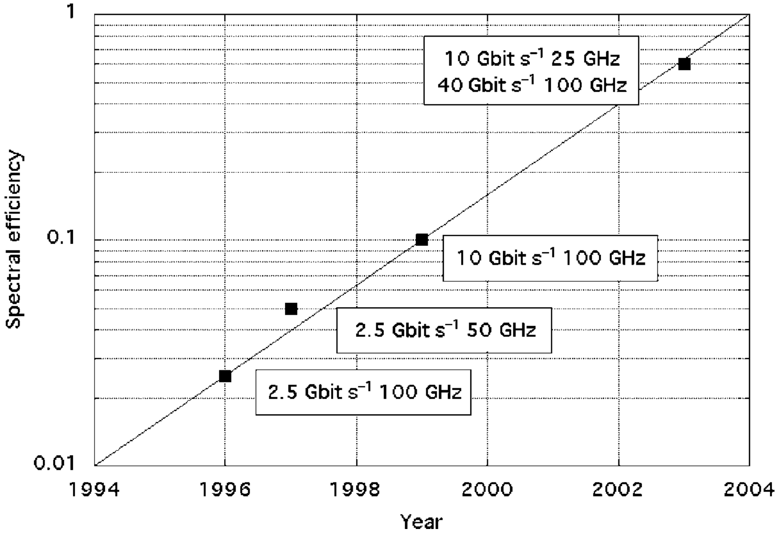


Figure C1.1.8. Historic evolution of the spectal efficiency (in bit s^{-1} Hz^{-1}) of optical systems.

drawback is that it does not allow to fully exploit the large available bandwidth of the fibre. As mentioned in the previous section, more sophisticated schemes can be used to improve the spectral efficiency, in particular multilevel modulation schemes. However, it is also well known that the modulation format has a strong influence on the penalties imposed by propagation phenomena. For example, to achieve the same BER, multilevel modulation schemes require a higher power than a binary modulation scheme, which means a higher sensitivity to propagation effects. Alternate schemes have been extensively investigated for long-haul systems, especially for undersea lightwave transmission. The duobinary code has been proposed by different laboratories as a good candidate, providing a better resistance to chromatic dispersion [60]; phase shaped binary transmission is another technique, which is a modified version of duobinary code [14]. Vestigial side band (VSB), a very well-known technique in radio communication systems, could also allow to reduce channel spacing and thus improve spectral efficiency [8, 9]. More recently, differential phase shift keying has also been proposed [24]. Although OOK remains today practically the only one used, at least for terrestrial systems, it is clear that improving the modulation format to better utilize the capacity of existing systems is a key issue in research. This effort is likely to lead to major improvements in the future optical transmission systems.

Selected recent results

Recently, publications show that the field of communication systems is very prolific, and novel systems with very impressive performance continue to be demonstrated. Table C1.1.1 summarizes the characteristics of some of the most remarkable systems reported in recent years. The objective is not to be exhaustive—dozens of examples could be compiled—but to illustrate some of the concepts mentioned earlier, and to give the reader a fair notion of the state of the art in research and development at the time of this writing. It has to be kept in mind that comparison of capacity is not necessarily meaningful, because transmission distance, amplification span and spectral efficiency are not identical.

Table C1.1.1. Selected examples of recent experimental high-capacity transmission lines.

Bit rate (Gbit s^{-1})	Number of channels	Total capacity (Tbit s^{-1})	Channel spacing (GHz)	Modulation scheme	Spectral efficiency (bit s^{-1} Hz^{-1})	Optical reach (km)	Amplification span (km)	Bands	Ref.
Terrestrial systems									
42.7	125	5		VSB		1200	100	C + L	[7]
40	80	3.2		Duobinary		300, G.655	100	C + Raman	[10]
40	273	10.9				116	58	S + C + L	[21]
		10.9		VSB + Pol. div. mult.	1.28	100	NA	Raman	[8]
Transoceanic systems									
365	10	3.65		NRZ + error correcting code		6850	~50		[57]
Repeaterless systems									
160	10	1.6	25			380	NA	Raman + remote C-band EDFA	[39]
104	40	4.16	125		0.32	135	NA	S + C + L + Raman	[11]
25	40	1				306	NA		[55]

References

[1] Agrawal G P 1995 *Nonlinear Fiber Optics* 2nd edn (San Diego, CA: Academic)

[2] Ait Sab O and Fang J 1999 Concatenated forward error correction schemes for long-haul DWDM optical transmission systems *Proc. of 25th European Conf. on Optical Communication, Nice, France (vol 2)* pp 290–291

[3] Andresciani D, Curti F, Matera F and Daino B 1987 Measurement of the group delay difference between the principal states of polarization on a low birefringence terrestrial fibre cable *Opt. Lett.* **12** 844–846

[4] Arbore M A, Zhou Y, Keaton G and Kane T J 2003 30dB gain at1500 nm in S-band erbium-doped silica fiber with distributed ASE suppression *Proc. of the SPIE—The International Society of Optical Engineering (vol 4989)*

[5] Bergh R A, Digonnet M J F, Lefevre H C, Newton S A and Shaw H J 1982 Single mode fiber optic components *SPIE Proceedings on Fiber Optics Technology'82 (vol 326)* pp 137–142

[6] Betti S, De Marchis G and Iannone E 1994 Toward an optimum use of the optical channel capacity *Fiber Integrated Opt.* **13** 147–164

[7] Bigo S *et al* 2001 Transmission of 125 WDM channels at 42.7 Gbit/s (5 Tbit/s capacity) over 12 × 100 km of TeraLight Ultra fibre *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 6)* pp 2–3

[8] Bigo S 2002 Improving spectral efficiency by ultranarrow optical filtering to achieve multiterabit/s capacities *Proc. of Opt. Fiber Commun. Conf., Anaheim, CA (vol 1)* pp 362–364

[9] Bigo S *et al* 2001 10.2 Tbit/s (256 × 42.7 Gbit/s PDM/WDM) transmission over 100 km TeraLight™ fiber with 1.28 bit/s/Hz spectral efficiency *Proc. of Opt. Fiber Commun. Conf., Anaheim, CA (Paper PD25-1-3, vol 4)*

[10] Bissessur H *et al* 2001 3.2 Tbit/s (80 × 40 Gbit/s) C-band transmission over 3 × 100 km with 0.8 bit/s/Hz efficiency *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 6)* pp 22–23

[11] Boubal F *et al* 2001 4.16 Tbit/s (104 × 40 Gbit/s) unrepeatered transmission over 135 km in S + C + L bands with 104 nm total bandwidth *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 1)* pp 58–59

[12] Brandon E and Blondel J P 1998 Raman limited, truly unrepeatered transmission at 2.5 Gbit/s over 453 km with 30 dBm launch signal power *Proc. of 24th European Conf. on Optical Communication, Madrid, Spain (vol 1)* pp 563–564

[13] Chan V W S 1997 Coding and error correction in optical fiber communications systems *Optical Fiber Communications III* vol A, ed I P Kaminow and T L Koch (New York: Academic)

[14] Charlet G *et al* 6.4 Tb/s (159 × 42.7 Gb/s) capacity over 21 × 100 km using bandwidth limited phase shaped binary transmission *ECOC 2002 Copenhague Post Deadline Paper 4.1*

[15] Chraplyvy A R 1990 Limitations on lightwave communications imposed by optical-fiber nonlinearities *J. Lightwave Technol.* **8** 1548–1557

[16] Ciprut P, Gisin B, Gisin N, Passy R, Von Der Weld P, Prieto F and Zimmer C W 1998 Second-order polarization mode dispersion: impact on analog and digital transmissions *J. Lightwave Technol.* **16** 757–771

[17] Desurvire E 1996 *Erbium-Doped Fiber Amplifiers: Principles and Applications* (New York: Wiley)

[18] Digonnet M J F 2001 *Rare Earth Doped Fiber Lasers and Amplifiers* 2nd edn (New York: Dekker)

[19] Eggleton B J *et al* 2000 Integrated tunable fiber gratings for dispersion management in high-bit rate systems *J. Lightwave Technol.* **18** 1418–1432

[20] Foschini G J and Poole C D 1991 Statistical theory of polarization mode dispersion in single mode fibers *J. Lightwave Technol.* **9** 1439–1456

[21] Fukuchi K *et al* 2001 10.92 Tbit/s (273 × 40 Gbit/s) triple-band/ultra-dense WDM optical-repeatered transmission experiment *Proc. of Opt. Fiber Commun. Conf., Anaheim, CA (Paper PD 24-1-3, vol 4)*

[22] Gisin N, Passy R, Bishoff J C and Perny B 1993 Experimental investigations of the statistical properties of polarization mode dispersion in single mode fibers *IEEE Photon. Technol. Lett.* **5** 819–821

[23] Gisin N *et al* 1995 Definition of polarization mode dispersion and first results of the COST 241 Round Robin measurements *Pure Appl. Opt.* **4** 511–522

[24] Griffin R A and Carter A C 2002 Optical differential quadrature phase-shift key (oDQPSK) for high capacity optical transmission *Proc. of Opt. Fiber Commun. Conf., Anaheim, CA (vol 1)* pp 367–368

[25] Hansen P B *et al* 1995 529 km unrepeatered transmission at 2.488 Gbit/s using dispersion compensation, forward error correction, and remote post- and pre-amplifiers pumped by diode-pumped Raman lasers *Electron. Lett.* **31** 1460

[26] Hasegawa A 2000 An historical review of application of optical solitons for high speed communications *Chaos* **10** 475–485

[27] Helard J-F, Bougeard S and Citerne J 1999 Forward error correction coding schemes for optic fiber cable systems at 10 Gbit/s *Proc. of 25th European Conf. on Optical Communication, Nice, France (vol 2)* pp 288–289

[28] Islam M N 2002 Raman amplifiers for telecommunications *IEEE J. Sel. Top. Quantum Electron.* **8** 548–559

[29] Joindot I *et al* 1996 Les Télécommunications par fibres optiques *Collection Technique et Scientifique des Télécommunications, Dunod, Paris, France*

[30] Joindot M *Internal France Telecom R&D Publication*

[31] Kaiser P 1986 Network architecture and systems technology for future broadband ISDN systems *Tech. Dig. of the 12th European Conf. on Optical Communication, ECOC'86, Barcelona, Spain*

[32] Kaminow I P and Koch T L 1997 *Optical Fiber Communications III* vols A and B (New York: Academic)

[33] Kawakami H, Miyamoto Y, Kataoka T and Hagimoto K 1994 Overmodulation of intensity modulated signals due to stimulated Brillouin scattering *Electron. Lett.* **30** 1507

[34] Kazovsky L, Benedetto S and Willner A 1996 *Optical Fiber Communication Systems* (Boston: Artech House)

[35] Kim K S, Stolen R H, Reed W A and Quoi K W 1994 Measurement of the nonlinear index of silica-core and dispersion-shifted fibers *Opt. Lett.* **19** 257–259

[36] Korotky S K, Hansen P B, Eskildsen L and Veselka J J 1995 Efficient phase modulation scheme for suppressing stimulated Brillouin scattering *Tech. Dig. of IOOC'95, Paper WD2-1, Hong Kong*

[37] Kuwano S, Yonenaga K and Iwashita K 1995 10 Gbit/s repeaterless transmission experiment of optical duobinary modulated signal *Electron. Lett.* **31** 1359

[38] Le Guen D, Lobo S, Merlaud F, Billes L and Georges T 25 GHz spacing DDWDM soliton transmission over 2000 km of SMF with 25 dB/span *ECOC 2001 Amsterdam Session WeF1*

[39] Le Roux P *et al* 2001 25 GHz spaced DWDM 160 × 10.66 Gbit/s (1.6 Tbit/s) unrepeatered transmission over 380 km *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 6)* pp 10–11

[40] Mao X P *et al* 1992 Stimulated Brillouin threshold dependence on fiber type and uniformity *IEEE Photon. Technol. Lett.* **4** 66

[41] Mears R J, Reekie L, Poole S B and Payne D N 1985 Neodymium-doped silica single-mode fibre lasers *Electron. Lett.* **21** 738–740

[42] Mears R J, Reekie L, Jauncey I M and Payne D N 1987 Low-noise erbium-doped fibre amplifier operating at 1.54 μm *Electron. Lett.* **23** 1026–1028

[43] Mollenauer L F, Evangelides S G Jr and Haus H A 1991 Long-distance soliton propagation using lumped amplifiers and dispersion shifted fiber *J. Lightwave Technol.* **9** 194–197

[44] Mollenauer L F and Mamyshev P V 1998 Massive wavelength-division multiplexing with solitons *IEEE J. Quantum Electron.* **34** 2089–2102

[45] Park Y K *et al* 1993 A 5 Gb/s repeaterless transmission system using erbium-doped fiber amplifiers *IEEE Photon. Technol. Lett.* **5** 79

[46] Penninckx D and Lanne S 2001 Reducing PMD impairments *Proc. of Opt. Fiber Commun. Conf., Anaheim, CA (Paper TuP1-1-4 vol 2)*

[47] Pureur D *Highwave Optical Technologies, France, Private Communication*

[48] Ragdale C M, Payne D N, De Fornel F and Mears R J 1983 Single-mode fused biconical taper fibre couplers *Proc. of the 1st International Conf. on Optical Fibre Sensors, London, UK* pp 75–78

[49] Sakamoto T, Aozasa S and Shimizu M Recent progress on S-band amplifiers *ECOC 2002 Copenhague Session Mo2.2*

[50] Segi T, Aizawa T, Sakai T and Wada A Silica-based composite fiber amplifier with 1480–1560 nm seamless gain band *ECOC 2001 Amsterdam Session MoL3*

[51] See, for example, Schicketanz D and Zeidler G 1975 GaAs-double-heterostructure lasers as optical amplifiers *IEEE J. Quantum Electron.* **11** 65–69 (and references therein)

[52] Shimizu M, Yamada M, Horiguchi M, Takeshita T and Okayasu M 1990 Erbium-doped fibre amplifiers with an extremely high gain coefficient of 11.0 dB/mW *Electron. Lett.* **26** 1641–1643

[53] Sunnerud H *et al* 2001 Long-term 160 Gb/s-TDM, RZ transmission with automatic PMD compensation and system monitoring using an optical sampling system *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 6)* pp 18–19

[54] Takushima Y and Okoshi T 1992 Suppression of stimulated Brillouin scattering using optical isolators *Electron. Lett.* **29** 1155

[55] Tanaka K, Sakata H, Miyakawa T, Morita I, Imai K and Edagawa N 2001 40 Gbit/s × 25 WDM 306 km unrepeatered transmission using 175 μm^2-Aeff fibre *Electron. Lett.* **37** 1354–1356

[56] Tkach R W *et al* 1995 Four-photon mixing and high-speed WDM systems *J. Lightwave Technol.* **13** 841–849

[57] Vareille G, Julien B, Pitel F and Marcerou J F 2001 3.65 Tbit/s (365 × 11.6 Gbit/s) transmission experiment over 6850 km using 22.2 GHz channel spacing in NRZ format *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 6)* pp 14–15

[58] Vengsarkar A M *et al* 1996 Long-period fiber gratings as band-rejection filters *J. Lightwave Technol.* **14** 58–64

[59] Wedding B, Chiarotto A, Kuebart W and Bulow H 2001 Fast adaptive control for electronic equalization of PMD *Proc. of Opt. Fiber Commun. Conf., Anaheim, CA (Paper TuP4-1-3, vol 2)*

[60] Yonenaga K and Kuwano S 1997 Dispersion-tolerant optical transmission system using duobinary transmitter and binary receiver *J. Lightwave Technol.* **15** 1530–1537

[61] Jianjun Yu *et al* 2001 160 Gb/s single-channel unrepeatered transmission over 200 km of non-zero dispersion shifted fiber *Proc. of 27th European Conf. on Optical Communication, Amsterdam, Netherlands (vol 6)* pp 20–21

C1.2

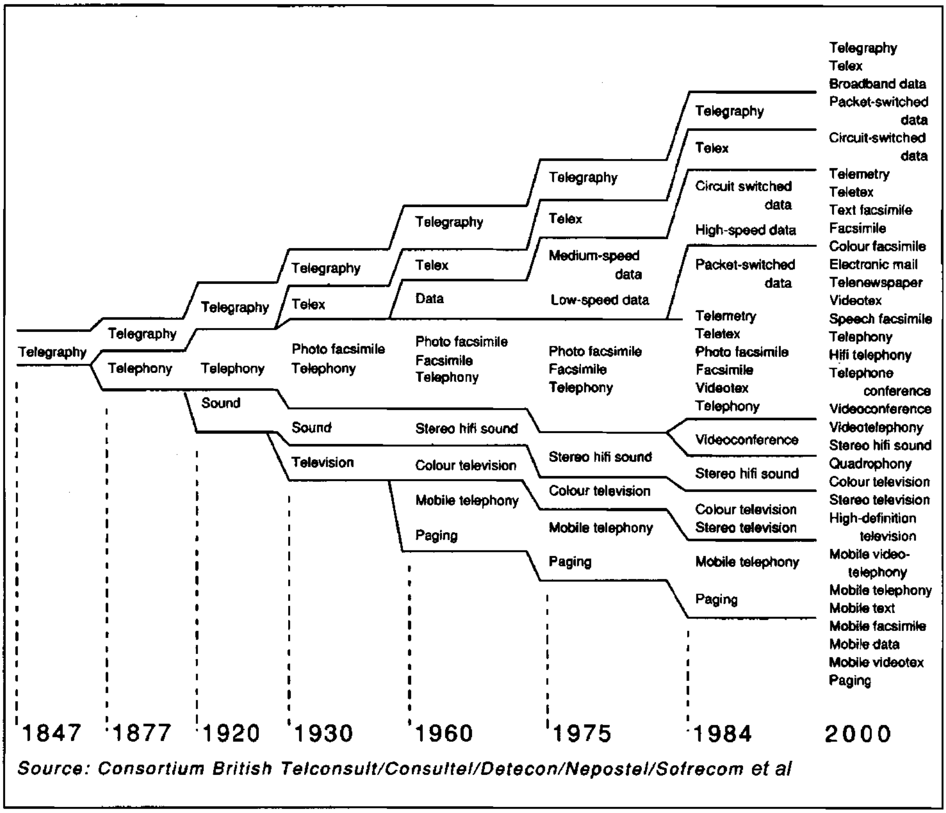
Optical network architectures

Ton Koonen

C1.2.1 Introduction

Telecommunication networks in all their various shapes are indispensable to bring information quickly anywhere and anytime, which is a vital need of our modern global society. Since the invention of the electrical telegraph by Samuel Morse in 1837, the variety of telecommunication services has grown at an increasing pace, as illustrated in figure C1.2.1. In addition, the services are becoming ever more individualized, and along with the penetration of video-based services ('a picture says more than a thousand words') the request for information transport capacity has exploded and is continuing to do so. Since the early 1990s, the introduction of the worldwide Internet has drastically promoted this information transport explosion. The number of Internet hosts is still increasing exponentially; from January 1992 to January 1997 to January 2002, it has grown from 727 thousand to 19.5 million to 147 million worldwide. This is causing data traffic to take an ever-larger share of the telecommunication network capacity; since a few years, it has surpassed the volume of the traditional voice traffic (but not yet its revenues). Wireless mobile telecommunication is attracting ever more users, and enables a fast roll-out of services to the end users without the need to install extensive first-mile customer access networks. The telecommunication market liberalization has provided ample opportunities to the entry of more operators and service providers, and the resulting national and international competition is pressing for very efficient high-capacity telecommunication networks.

As a result, the volume of telecommunication traffic is ramping up at a compound annual growth rate of roughly 60%, which means an increase with a factor of 10 in no more than 5 years. This traffic is in vast majority carried along fixed-wired networks, due to their high reliability, security, and immunity for external disturbances. Wireless networks are coming up in customer access environments; but due to increasing microwave carrier frequencies and user densities, the wireless cells are shrinking and thus extensive fixed access network lines are still indispensable as the vessels to feed the wireless antenna stations. Traditional coaxial and twisted-pair copper cables have been the transport media of choice since the introduction of telecommunication networks. However, the advent of optical fibre with its extremely low losses and extremely large bandwidth as pioneered by Kao and Hockam in 1966, and the commercial introduction of optical fibre communication systems in the early 1980s has caused that single-mode optical fibre has become the transport medium uniquely used in long-distance fixed-wired core transport networks, and that it is also conquering at increasing pace the area of metropolitan and access networks.


Telegraphy
1847
Telegraphy
Telephony
1877
Telegraphy
Telephony
Sound
1920
Telegraphy
Telex
Photo facsimile
Telephony
Sound
Television
1930
Telegraphy
Telex
Data
Photo facsimile
Facsimile
Telephony
Stereo hifi sound
Colour television
Mobile telephony
Paging
1960
Telegraphy
Telex
Medium-speed data
Low-speed data
Photo facsimile
Facsimile
Telephony
Stereo hifi sound
Colour television
Mobile telephony
Paging
1975
Telegraphy
Telex
Circuit switched data
High-speed data
Packet-switched data
Telemetry
Teletex
Photo facsimile
Facsimile
Videotex
Telephony
Videoconference
Stereo hifi sound
Colour television
Stereo television
Mobile telephony
Paging
1984
Telegraphy
Telex
Broadband data
Packet-switched data
Circuit-switched data
Telemetry
Teletex
Text facsimile
Facsimile
Colour facsimile
Electronic mail
Telenewspaper
Videotex
Speech facsimile
Telephony
Hifi telephony
Telephone conference
Videoconference
Videotelephony
Stereo hifi sound
Quadrophony
Colour television
Stereo television
High-definition television
Mobile video-telephony
Mobile telephony
Mobile text
Mobile facsimile
Mobile data
Mobile videotex
Paging
2000

Source: Consortium British Telconsult/Consultel/Detecon/Nepostel/Sofrecom et al

Figure C1.2.1. Evolution of telecommunication services.

C1.2.2 Telecommunication networks hierarchy

Telecommunication networks are carrying traffic at various aggregation levels, as illustrated in figure C1.2.2.

At the highest level, *long-haul core networks* are transporting huge data capacities in the tens of terabits/second over large distances, such as over transnational and transoceanic links up to 9000 km (transpacific). These global and wide area networks are transporting circuit-switched data following the SDH (or SONET) standards, where each fibre usually carries a number of wavelengths at bitrates of up to 10–40 $\mathrm{Gbit\,s^{-1}}$ each. They have to meet extremely high levels of reliability and availability, considering the huge volume of customers dependent on them. Taking into account the increasing dynamics in the traffic matrix describing the data flows between the various network nodes, packet switching techniques are being introduced, which can offer a more efficient utilization of the network's resources than circuit switching.

Metropolitan area networks (MANs) are covering large urban areas with a reach of up to 100 km and capacities of tens of gigabit/second, serving in particular business parks and residential customer access regions. High availability for large-volume fast file transfer is a major need for the business customers. Also these networks should be easily scalable for adding more network nodes, and flexible to accommodate new business needs. Storage area networks (SANs) are specifically employed for regularly moving large volumes of data between geographically separated sites, in order to safeguard vital business information.

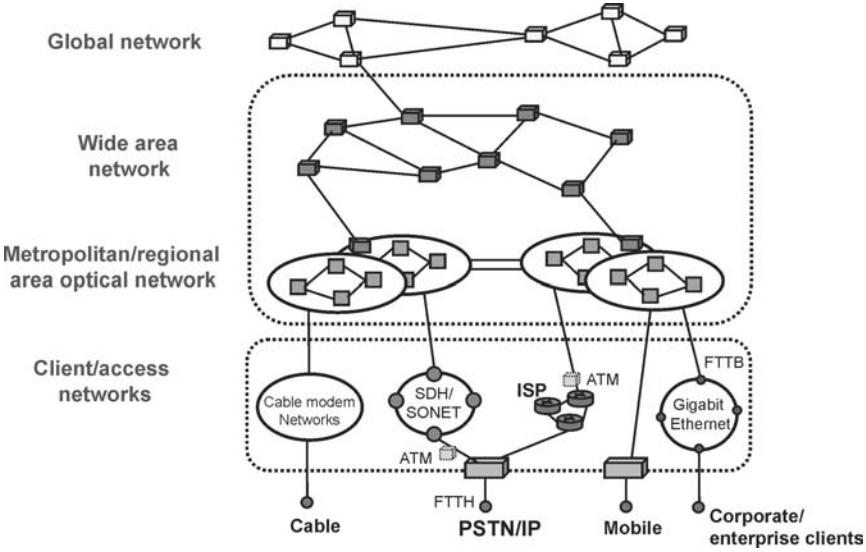


Figure C1.2.2. The hierarchy of telecommunication networks.

Access networks are providing a wide variety of services to the end customers, and consist of fibre feeder networks followed by various first-mile networks. These networks are mostly optimized for a particular set of services, and exploited by different operators. Coaxial cable network operators offer television and radio broadcast services, and since recently also data modem services and telephony, multiplexed in different frequency bands. The public switched telephone network (PSTN) uses twisted pair copper cables and is carrying voice telephony and data services, time-multiplexed according to the SDH/SONET or ATM standard; it is exploited by the incumbent telecom operators as well as new entrants. Mobile network operators are mainly providing wireless voice telephony, according to among others the GSM standard; also wireless data services using GPRS and UMTS are coming up. The statistics of traffic in access networks shows much higher dynamics than in metropolitan and core networks, due to the significantly lower traffic aggregation levels. Therefore, applying packet switching instead of circuit switching can improve remarkably the network utilization efficiency. Access networks have to be laid out very cost-effectively, as the factor with which network equipment is shared among customers is much lower than in metropolitan and in core networks.

C1.2.2.1 Network topologies

A number of general network topologies as shown in figure C1.2.3 can be discerned for implementation of the various hierarchical network layers. Each of them has a specific set of characteristics, which makes it suited to match the requirements of a certain layer. *Mesh networks*, as exemplified in figure C1.2.3(*a*), provide a number of options to route traffic between two network nodes. This routing redundancy yields a large availability of the network services, which is a highly valued merit in long-distance core and metro-core networks. The entailed extra costs are of less concern due to the large resource-sharing factor among the huge customer base served. *Ring networks*, in particular when composed of both an inner and an outer ringlet as shown in figure C1.2.3(*b*), provide clockwise and counterclockwise traffic routing options; thus also network protection is established in order to yield a good network availability (albeit at a lower level than in mesh networks, but also at lower costs as less network resources

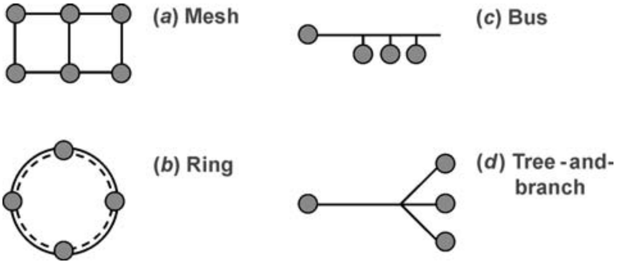


Figure C1.2.3. Network topologies.

are needed). The combination of good availability and moderate costs of resources makes ring topologies well suited to implement metropolitan networks. *Bus networks*, as shown in figure C1.2.3(*c*), use a common single linear medium along which signal power is tapped off to the various nodes (with power losses accumulating at each subsequent tap). This topology is quite cost-efficient, as a minimum of network resources is needed. The nodes may exchange information as peers in the network, which makes this topology suited for linking data processing equipment. However, no routing redundancy is provided, and therefore there is no guarantee for good network availability. In *tree-and-branch networks*, exemplified in figure C1.2.3(*d*), a relatively long single feeder line is running from a headend node to a power splitting point, from where the signal power is distributed via short lines to a number of end nodes. This topology is most suited for broadcasting information from a single headend node to many customer end nodes. No routing redundancy is provided, so again network availability is limited. Also the topology is very cost-efficient, as the costs of the feeder line and the headend equipment is shared by all the end nodes. In addition, the signal power distribution is more efficient than in the bus network, as the power splitting loss (in decibels) to an end node increases only with the logarithm of the number of end nodes whereas it increases linearly with this number in the bus network. Every end node receives the same power level, which relaxes the dynamic range over which the end node equipment has to operate. The tree-and-branch topology is therefore well suited and popular for access networks. With a fully passive optical power splitter in the branching point, the topology is also widely known as the passive optical network (PON).

In the next sections of this chapter, architectural aspects and key functionalities needed at the subsequent hierarchical network layers (core networks, metropolitan networks, and access networks) will be discussed in more detail.

C1.2.3 Core networks

The main task of core networks is to transport huge amounts of telecommunication traffic over large distances in a highly reliable way. The network should not break down by failures in one or in a few links. Therefore, provisions have to be included for alternative routing of traffic, which are adequately offered by the mesh network topology with crossconnect functions in the nodes. The links between the nodes are usually very long (>100 km), and intermediate signal amplification and compensation of fibre dispersion effects is needed.

C1.2.3.1 Optical signal multiplexing techniques

Using *electrical time division multiplexing* (ETDM), commercial systems support bitrates up to 10–40 Gbit s^{-1}, and by direct laser diode modulation (or a laser diode followed by an external

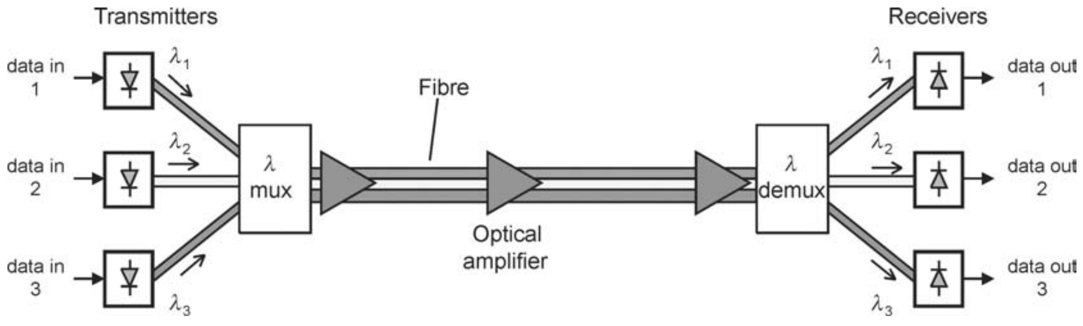

Transmitters
Receivers
data in 1
data in 2
data in 3
λ_1
λ_2
λ_3
λ mux
Fibre
Optical amplifier
λ demux
data out 1
data out 2
data out 3

Figure C1.2.4. Wavelength division multiplexing.

modulator) these bitrates can be carried through a fibre link via single wavelength channel. The transport capacity of the fibre link can basically be increased further in two ways: by increasing the number of wavelength channels, and by increasing the bitrate per channel.

Multiple wavelength channels can be carried by a single fibre by combining them at the transmitting end by means of a wavelength multiplexing device, and separating them again at the receiving end by a wavelength demultiplexer. This *wavelength division multiplexing* (WDM) approach is shown in figure C1.2.4. Standard single-mode fibre offers low dispersion in the wavelength window of 1285–1330 nm, which amounts to about 8 THz of bandwidth, and low attenuation in the 1500–1580 nm window, corresponding to about 10 THz. The popular 1530–1560 nm window (the C-band, the operation range of erbium-doped fibre optical amplifiers) represents about 3.8 THz of bandwidth. Wavelength channel spacings of 100 GHz according to ITU-T G.692 (or even down to 25 GHz) are being deployed, which enables hundreds of wavelength channels to be accommodated in a single fibre. Commercial systems are available which carry 160 wavelength channels at 10 Gbit s^{-1} each, amounting to a total 1.6 Tbit s^{-1} capacity. The record obtained in research stands at 273 wavelengths at 40 Gbit s^{-1} each, amounting to 10.92 Tbit s^{-1} [1].

Increasing the bitrate per wavelength channel beyond the limits of ETDM can be achieved with the so-called *optical time division multiplexing* (OTDM) technique. As illustrated in figure C1.2.5, at the transmitter side a single laser diode generates a sequence of equidistant narrow optical pulses. After splitting and distribution to four fast optical gates, the pulse train is on/off switched in each gate by an

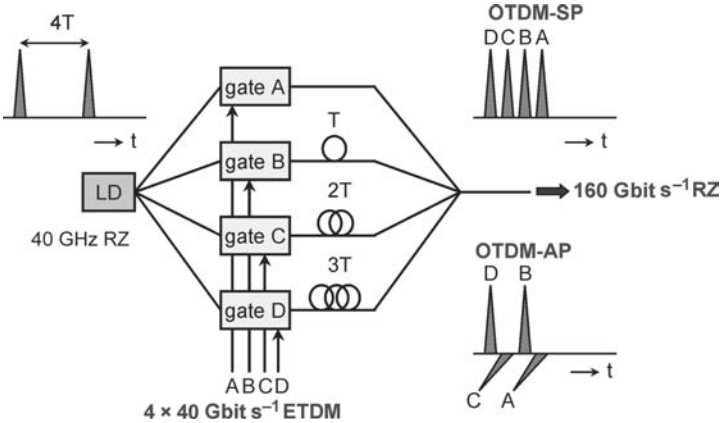

4T
→ t
LD
40 GHz RZ
gate A
gate B
gate C
gate D
T
2T
3T
A B C D
4 × 40 Gbit s⁻¹ ETDM
OTDM-SP
D C B A
→ t
160 Gbit s⁻¹ RZ
OTDM-AP
D B
C A
→ t

Figure C1.2.5. Optical time division multiplexing transmitter.

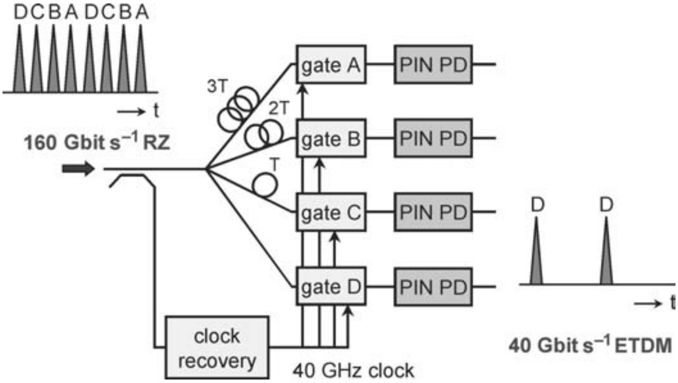


Figure C1.2.6. Optical time division demultiplexing receiver.

electrical time-multiplexed data stream. The modulated pulse trains are delayed with respect to each other, and subsequently interleaved (like a 'zipper'). The individual pulse trains may all have the same polarization (SP, single polarization), or alternating polarization (AP). The optical pulses need to be sufficiently narrow in order to avoid crosstalk. With the OTDM-AP scheme, somewhat broader optical pulses are allowed than with the OTDM-SP scheme. The resulting output signal is a modulated optical pulse data stream at a speed that is the sum of the speeds of the electrical input data streams. At the receiver side, first the clock signal needs to be recovered from the high-speed data stream. Using this clock signal and appropriate time delays, fast optical gates followed by optical receivers with PIN photodiodes can demultiplex the original constituting pulse trains, as shown in figure C1.2.6. Using these OTDM techniques, the next bitrate hierarchy of 160 Gbit s^{-1} can be realized with 40 GHz electronics. In research, the record has been set at 1.28 Tbit s^{-1} by polarization-multiplexing two OTDM-SP 640 Gbit s^{-1} streams [2].

The chart shown in figure C1.2.7 indicates how by increasing the data rate per wavelength channel by means of advances in electrical and optical TDM at one hand, and by increasing the number of

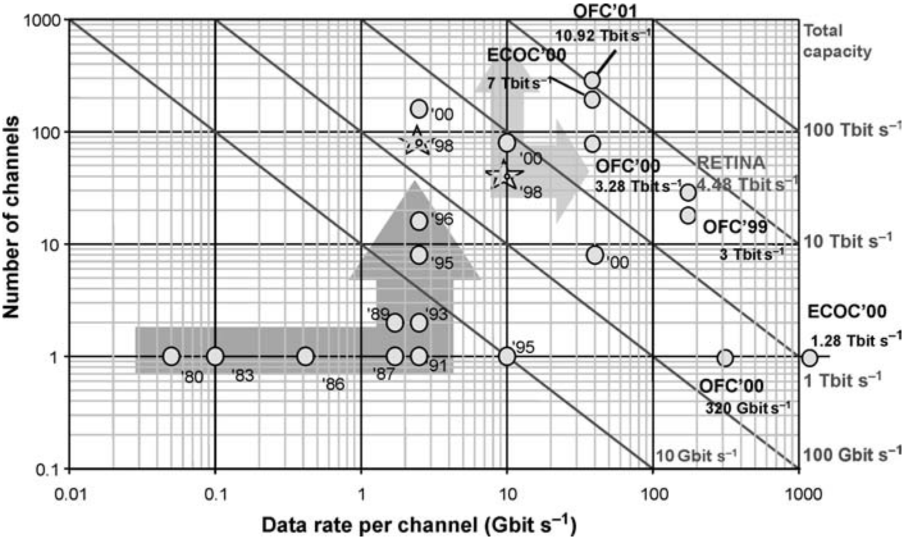


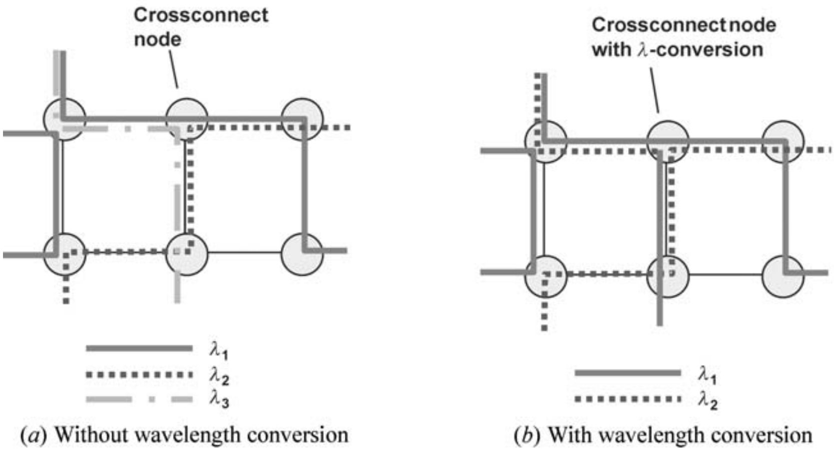
Figure C1.2.7. The evolution of transport capacity of a single fibre link.

wavelength channels at the other hand, the transport capacity of a single fibre has been enormously increased since the introduction of optical fibre communication systems in the early 1980s. The main leap forward was made with the introduction of wavelength multiplexing in the early 1990s. And opportunities for further capacity growth are still being created by opening new wavelength bands such as the S-band from 1450 to 1530 nm, and the L-band from 1560 to 1620 nm, supported by alternative optical amplifying processes due to fibre nonlinearities such as Raman gain. Optical gain in the 1300 nm window can be provided by fibre amplifiers using rare earth materials such as praseodymium and neodymium, and by semiconductor optical amplifiers. The present system capacity record stands at 10.92 $\mathrm{Tbit\,s^{-1}}$, deploying 273 wavelength channels at 40 $\mathrm{Gbit\,s^{-1}}$ each, as mentioned earlier. Applications for multi-terabits/second systems are found not only in transcontinental and transoceanic systems, but also for massive data processing systems such as huge synthesized antenna array systems for astronomical observations (e.g. the RETINA system [3]).

In addition to wavelength multiplexing, optical amplification and dispersion compensation are key techniques to enable high volume data transport over long fibre links. A single optical amplifier can handle many wavelength channels simultaneously; otherwise, an equivalent number of opto-electronic regenerators plus a wavelength multiplexer and demultiplexer would be needed, which is clearly more costly, requires more maintenance and powering, and requires a sizeable adaptation effort when the system needs to be upgraded with more wavelength channels.

C1.2.3.2 Traffic routing

The deployment of multiple wavelength channels is not only beneficial for increasing the transport capacity of a single fibre link, but also provides more flexibility to route traffic streams in the network. Wavelength channels can constitute independent optical paths through the network, and each path may be laid out individually to establish an optimum connection between certain end nodes. As shown in figure C1.2.8(*a*), within each node the wavelength channels are routed by means of optical crossconnects which guide the incoming signals depending on their wavelength and the entrance port to a specific output port. The routing table is usually set by the network management system, and can be altered when needed by changing traffic conditions. When two optical paths do not touch each other, they may be established with the same wavelength. This wavelength re-use reduces the overall number of



(*a*) Without wavelength conversion (*b*) With wavelength conversion

Figure C1.2.8. Establishing wavelength paths.

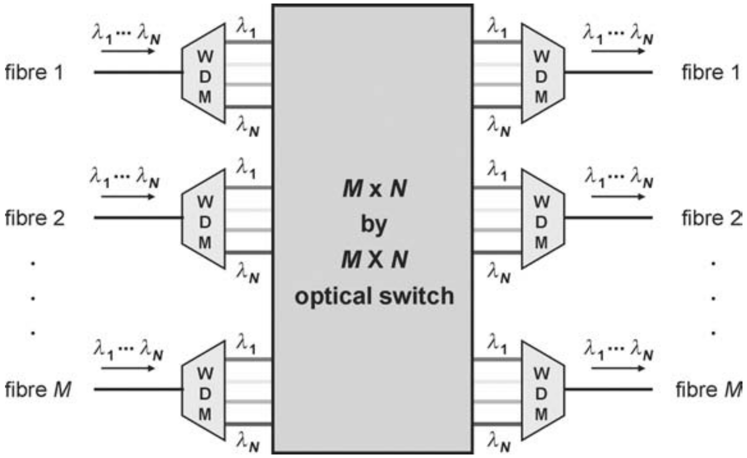


Figure C1.2.9. Optical crossconnect.

wavelengths needed in the network. In figure C1.2.8(*a*), for instance, wavelength λ_1 is used twice, but in different paths. As illustrated in figure C1.2.8(*b*), further reduction of the number of wavelengths needed can be obtained by using wavelength converters in the nodes; thus an optical path may be constituted by a sequence of different wavelengths in a series of links. By controlling the crossconnects in the nodes, the network management system can optimize the traffic flow routings to obtain a good balancing of the network load among the links, and thus to reduce the probability of congestion and to increase the network's efficiency. Also link failures can be circumvented by routing traffic through alternative links, which improves the reliability of the network.

The basic layout of an *optical crossconnecting node* is shown in figure C1.2.9. The N wavelength channels carried by each of the M input fibres are firstly separated by a wavelength demultiplexer, and subsequently a large optical matrix switch with $(M \times N)$ input ports and $(M \times N)$ output ports is needed which can route any wavelength channel from any input fibre to any of the M output fibres. For larger numbers M and N, the internal architecture of the matrix switch becomes quite comprehensive. Composed of individual 2×2 optical switches, a Benes switch architecture as shown in figure C1.2.10 would require $(MN/2)(2\,{}^2\log MN - 1)$ switches; and $2\,{}^2\log MN - 1$ switches have to be passed from any input to any output causing losses which increase with the switch size. Three-dimensional free-space optical switches using micro-mechanical mirrors for beam steering between the array of input fibres

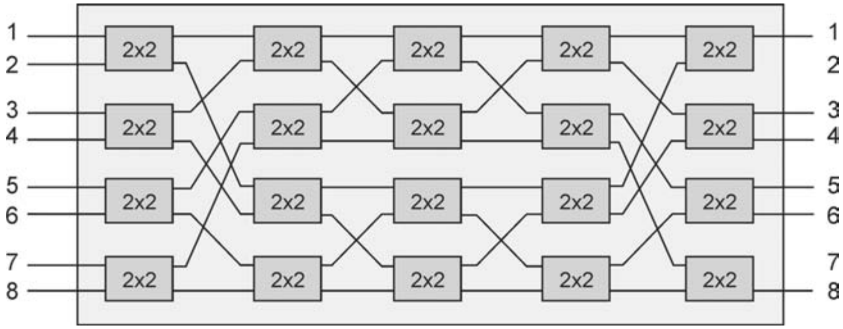


Figure C1.2.10. Re-arrangeable nonblocking Benes switching matrix.

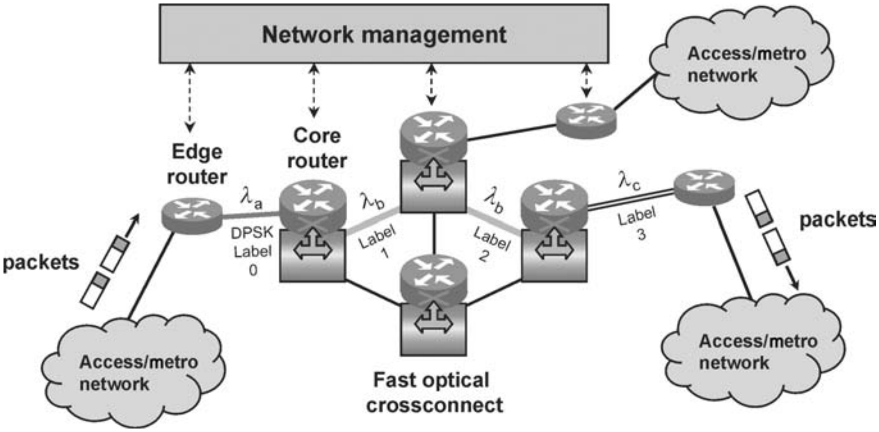


Figure C1.2.11. Optical packet routing.

and the array of output fibres exhibit losses which do not increase with the number of switch ports, and thus can outperform the Benes planar architecture [20, 21].

The amount of packet-based data traffic is growing fast in all telecommunication network layers, a.o. due to the steeply rising use of Internet services. IP packets are usually carried over SDH/SONET and/or ATM, which in their turn are carried in WDM channels; efficiency can be gained, however, by transporting the IP packets directly in the WDM channels, with some simple framing for basic synchronization functions. By using packet switching in the optical crossconnect nodes instead of circuit switching, the network capacity can be exploited much more dynamically in response to instantaneous traffic demands, and thus the network operation efficiency is improved. Following the GMPLS protocol, wavelength-switched paths can be set up in the network in a similar way as label-switched paths in the MPLS protocol. The packets can be marked by assigning a wavelength to them, which acts as a label. Based on these labels, a path is established for each packet through the network; see figure C1.2.11. The per packet label swapping needed for efficient forwarding and routing is achieved by wavelength converters in the nodes, which may be realized with fast tunable lasers and an all-optical wavelength converter (such as a Mach–Zehnder interferometer with semiconductor optical amplifiers in its branches). Another way to attach label information to a packet is to incorporate it in the packet data frame, or to modulate it on a subcarrier frequency outside the spectrum of the packet data. Even more comprehensive label information may be attached by frequency shift keying (FSK) modulation of the optical carrier (or differential phase shift keying, DPSK), orthogonally to the payload data that is intensity modulated on the carrier [4, 5]. Using wavelength converters and a wavelength-selective passive router (such as an arrayed waveguide grating router, AWGR), a fast optical crossconnect can be realized; e.g. in an architecture as shown in figure C1.2.12. Variable delay lines are included to avoid collisions of packets that are heading for the same output fibre at the same wavelength. Congestion can also be avoided by temporary buffering in the recirculating loops; the loops plus power splitter enable optical multicasting as well.

C1.2.4 Metropolitan area networks

MANs have to bring a variety of services to large urban areas, typically at bitrates up to 10 Gbit s^{-1} per wavelength over distances between nodes of less than 100 km. Major customers to be served are business parks, where customers mainly ask for fast large-volume file transfer, e.g. to interconnect their offices

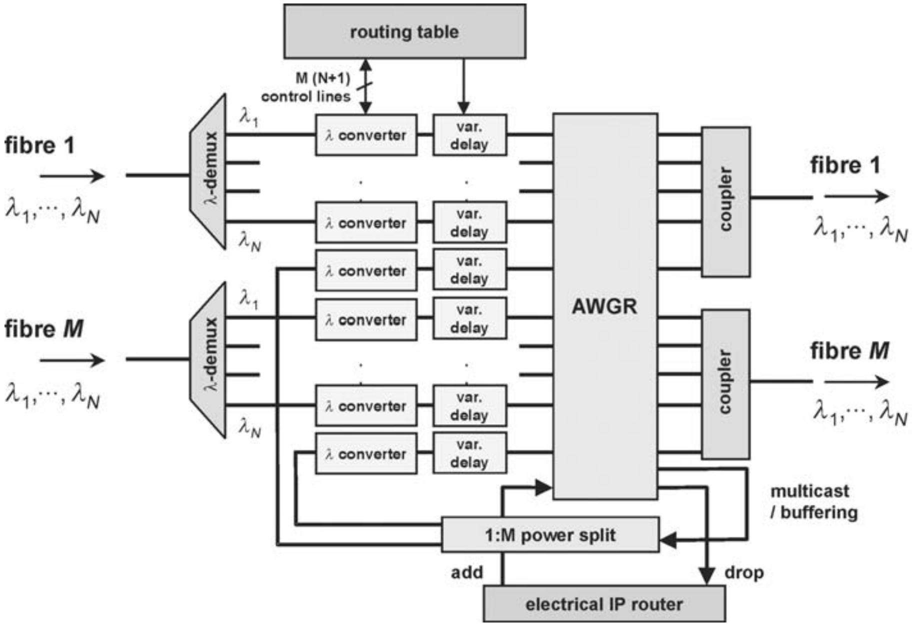

routing table
M (N+1) control lines
fibre 1
$\lambda_1,\cdots,\lambda_N$
λ-demux
λ_1
λ_N
λ converter
var. delay
fibre M
AWGR
coupler
fibre 1
$\lambda_1,\cdots,\lambda_N$
fibre M
$\lambda_1,\cdots,\lambda_N$
multicast / buffering
1:M power split
add
drop
electrical IP router

Figure C1.2.12. Fast optical crossconnect using wavelength conversion.

and for storage at safe locations of vital business data (storage area network). As high reliability is required while network costs per customer should be kept limited, the ring network topology is well suited for providing good network protection at moderate costs.

An example layout of a MAN ring network is shown in figure C1.2.13. Typically, the ring is composed of two fibres, on which the traffic flows clockwise and counterclockwise, respectively. The fibre links between neighbouring nodes are less than 20 km, and the ring circumference is usually less than 100 km; therefore no in-line optical amplifiers are needed. Each node uses a specific set of wavelength channels, typically carrying bitrates up to 2.5 Gbit s^{-1} per channel. Each node is communicating with the hub node; in the hub, opto-electric-optical translation to another wavelength set is done in order to establish communication between ring nodes. So virtually in the network there are point-to-point node-to-hub connections. The MAN typically contains up to 16 nodes, and up to 40 wavelength channels. Per node, one or more out of the set of wavelength channels can be dropped and/or added by means of optical add/drop multiplexers (OADMs). Through the hub, also communication to other networks can be established.

C1.2.4.1 Network protection

The two-fibre topology enables self-healing of the ring network in case of a cable break. As shown in figure C1.2.14, a link failure may be circumvented by looping back the traffic at the nodes neighbouring the broken fibre cable. This procedure also allows a ring segment to be taken temporarily out of service without disrupting the traffic on the remainder of the ring network, e.g. for maintenance or for inserting a new node. In the SONET standard, the 2-fibre uni-directional path-switched ring (UPSR) is composed of an outer primary path ring, carrying clockwise the normal working traffic; see figure C1.2.15. The inner ring provides the counter-clockwise protection path. In the SDH standard, this concept is called sub-network connection protection (SNCP). Both rings are fed from transmitters at the nodes, and at

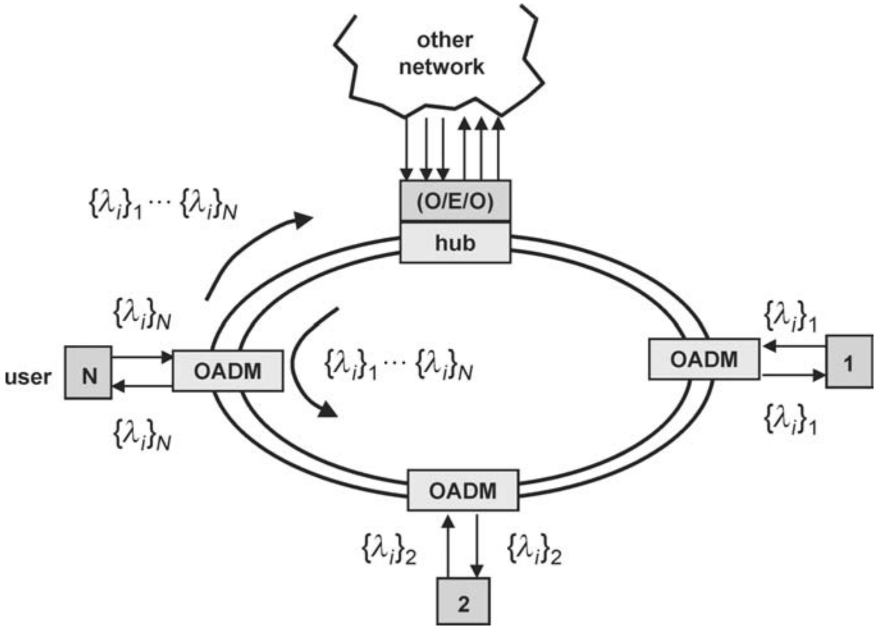

other network
(O/E/O)
hub
$\{\lambda_i\}_1 \cdots \{\lambda_i\}_N$
$\{\lambda_i\}_N$
user
N
OADM
$\{\lambda_i\}_1 \cdots \{\lambda_i\}_N$
OADM
$\{\lambda_i\}_1$
1
$\{\lambda_i\}_1$
$\{\lambda_i\}_N$
OADM
$\{\lambda_i\}_2$
$\{\lambda_i\}_2$
2

Figure C1.2.13. Multiwavelength MAN with fixed wavelength routing via hub.

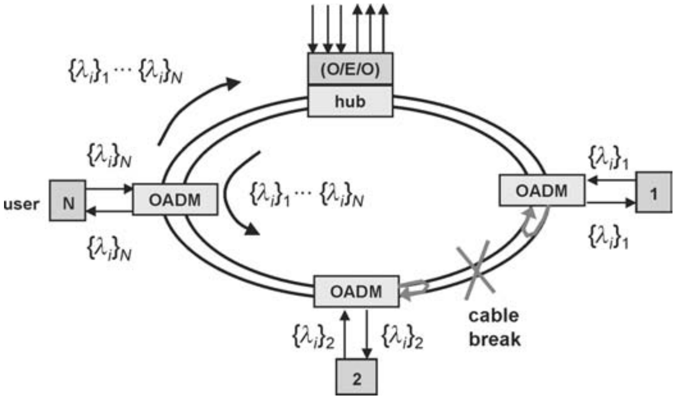

(O/E/O)
hub
$\{\lambda_i\}_1 \cdots \{\lambda_i\}_N$
$\{\lambda_i\}_N$
user
N
OADM
$\{\lambda_i\}_1 \cdots \{\lambda_i\}_N$
OADM
$\{\lambda_i\}_1$
1
$\{\lambda_i\}_1$
$\{\lambda_i\}_N$
OADM
cable break
$\{\lambda_i\}_2$
$\{\lambda_i\}_2$
2

Figure C1.2.14. Self-healing by looping back at the OADMs.

the node receiver, the signal from the ring with the best quality is chosen. Bi-directional traffic between two nodes will span around the entire ring, and thus consumes resources on every link of the ring. The 4-fibre bi-directional line-switched ring (BLSR) in the SONET standard (called multiplex section shared protection ring, MS-SPRing, in the SDH standard) consists of an outer bi-directional 2-fibre primary loop carrying the normal working traffic, and an inner bi-directional 2-fibre secondary loop for protection. Bi-directional traffic between two nodes is only sent along part of the ring, and does not involve resources on other parts of the ring. As shown in figure C1.2.16, the system is well protected against cable breaks or node failures. In IEEE 802.17, the resilient packet ring (RPR) is being discussed for standardization [6]. It consists of a dual counterpropagating ring, with up to 256 node stations

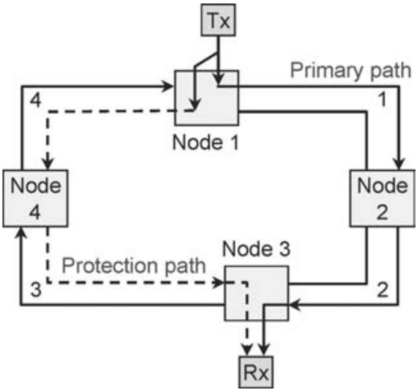
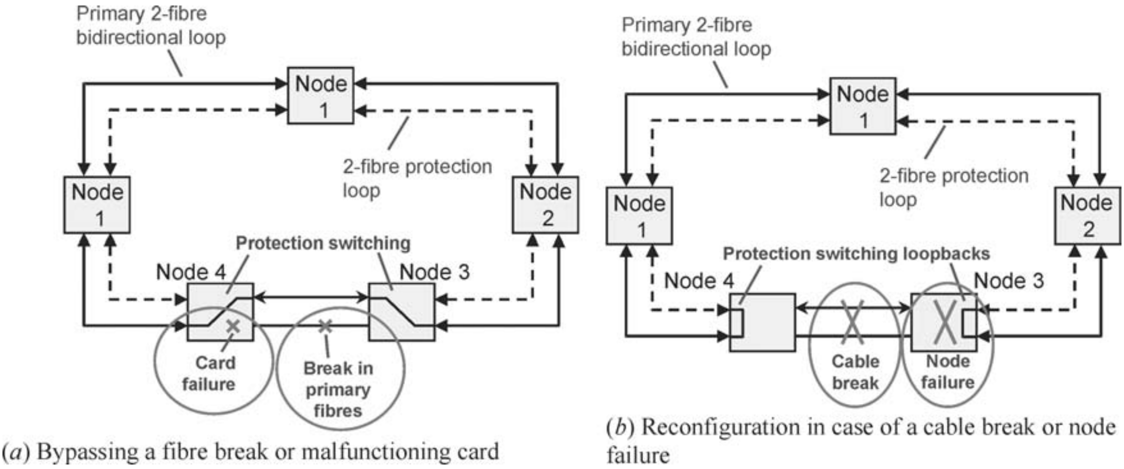

Tx
Primary path
1
4
Node 1
Node 4
Node 2
Node 3
Protection path
3
2
Rx

Figure C1.2.15. Protection in the 2-fibre SONET UPSR (SDH SNCP) network.


Primary 2-fibre bidirectional loop
Node 1
2-fibre protection loop
Node 1
Node 2
Protection switching
Node 4
Node 3
Card failure
Break in primary fibres
Primary 2-fibre bidirectional loop
Node 1
2-fibre protection loop
Node 1
Node 2
Protection switching loopbacks
Node 4
Node 3
Cable break
Node failure

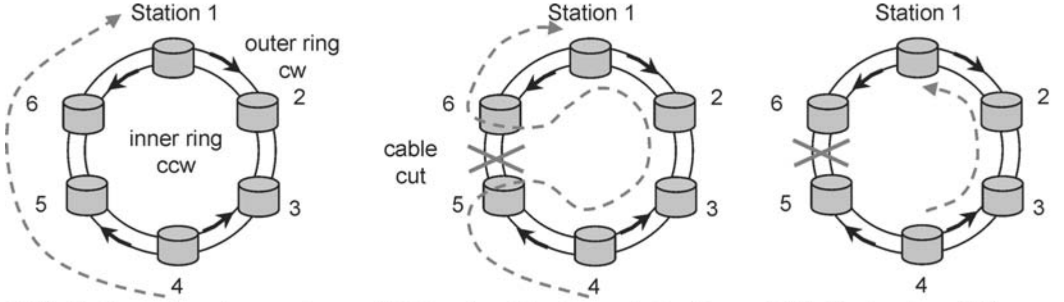
(*a*) Bypassing a fibre break or malfunctioning card

(*b*) Reconfiguration in case of a cable break or node failure

Figure C1.2.16. Protection in the 4-fibre SONET BLSR (SDH MS-SPRing) network.

and spans up to 6000 km. The RPR is able to operate with any packet protocol (such as Ethernet), and uses frame-based transmission based on the standard gigabit Ethernet frame of variable length, but also supporting jumbo frames up to 9 kbytes. It is optimized for providing multiple IP service classes: reserved traffic class A0 (guaranteed data rate, small bounded delay and jitter), high priority class A1 (committed rate, bounded delay and jitter), medium priority class B (committed rate, and excessive rate subject to fair access), and low priority class C (best effort, subject to fair access). The RPR management system is able to provide ring survivability within 50 ms. It selects the inner or outer ring to establish the shortest path. When a ring failure occurs, the traffic is first wrapped at the failure point, subsequently the damaged ring topology is discovered, and the source traffic is steered to establish the shortest path again taking the failure into account (see figure C1.2.17).

The nodes in the MAN may also route the wavelength channels dynamically between themselves without relaying via a hub, as exemplified in figure C1.2.18. By means of an appropriate algorithm, and by deploying tunable laser diodes and tunable OADMs at the nodes, the most appropriate wavelengths


Station 1
outer ring
cw
inner ring
ccw
6
2
5
3
4
Station 1
cable
cut
6
2
5
3
4
Station 1
6
2
5
3
4

(*a*) Traffic flows along the outer ring before cable cut

(*b*) Wrapping of the data path after the cable cut

(*c*) Traffic steering, yielding a shortest data path after discovery of the new topology

Figure C1.2.17. Protection in the resilient packet ring network.

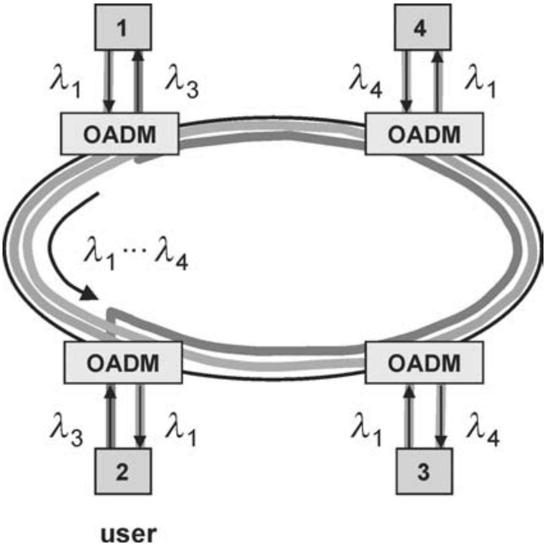

1
4
λ_1
λ_3
λ_4
λ_1
OADM
OADM
$\lambda_1 \cdots \lambda_4$
OADM
OADM
λ_3
λ_1
λ_1
λ_4
2
3
user

Figure C1.2.18. Multiwavelength MAN with dynamic wavelength routing between nodes.

can be assigned to the communication paths needed between nodes. When paths do not overlap, the same wavelength may be used (wavelength re-use), which reduces the number of wavelengths needed.

C1.2.4.2 Optical add/drop multiplexing

Essential network elements in the ring-shaped MANs are OADMs, which are able to extract data stream on a particular wavelength channel (or several channels) from the ring, and to insert new data streams on one or more wavelength channels into the ring. Signalling information controls whether a data stream will be dropped at a node or will pass through it. There are various ways to transfer this signalling information to the nodes, without interfering with the data streams. It may be modulated on a subcarrier frequency that is positioned above the data spectrum, and each wavelength channel may carry a unique subcarrier frequency. Thus, it is not needed to wavelength-demultiplex the channels first in order to detect and demodulate the signalling information on the various subcarriers. Another method to

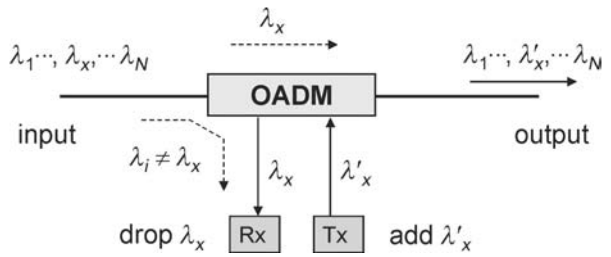

λ_x
$\lambda_1 \cdots, \lambda_x, \cdots \lambda_N$
OADM
$\lambda_1 \cdots, \lambda'_x, \cdots \lambda_N$
input
output
$\lambda_i \neq \lambda_x$
λ_x
λ'_x
drop λ_x
Rx
Tx
add λ'_x

Figure C1.2.19. Crosstalk in an Optical Add/Drop Multiplexer (OADM).

transfer the signalling information is to modulate it on a dedicated separate common wavelength channel, which is opto-electric-optically (O/E/O) converted at each node for inspection. And a third method is to put the signalling information in-band in a digital frame together with the data (e.g. the digital wrapper concept); then at each node all wavelength channels need to be demultiplexed and O/E/O converted for inspection.

The OADMs need to be wavelength-selective. When the pass-through wavelength channels experience some filtering, the passband of many OADMs put in cascade along the ring may narrow significantly. Thus the possibility to extend the number of nodes in the network scalability may be reduced. This scalability issue is avoided by applying notch-type OADMs, which do not exhibit pass-through filtering.

Another issue in the design of OADMs is crosstalk. Two types of crosstalk may be discerned: incoherent crosstalk and coherent crosstalk, as shown in figure C1.2.19. Incoherent crosstalk may occur because part of the wavelength channel(s) $\lambda_i \neq \lambda_x$ to be passed through are dropped, and interfere with the intentionally dropped wavelength channel λ_x. Coherent crosstalk occurs because part of the dropped wavelength channel λ_x leaks through, and beats with the added wavelength channel $\lambda_x{}'$ which has nominally the same wavelength. Thus the coherent crosstalk cannot be removed by subsequent optical bandpass filtering, and it may accumulate when cascading OADM nodes. The coherent crosstalk imposes the most stringent requirements on the device crosstalk characteristics. Mathematical analysis taking the specific statistical properties of the beat signal into account show that per node the crosstalk attenuation needs to be better than 32 dB to yield a bit error rate (BER) better than 10^{-12} [7]. For incoherent crosstalk in a four-channel OADM, the crosstalk attenuation needed to yield a BER $< 10^{-12}$ needs to be better than 13 dB only.

An example of an OADM that can drop and add a fixed wavelength channel is shown in figure C1.2.20. A fibre Bragg grating (FBG) reflects only wavelength λ_x; the other wavelength channels are passed unaffected. This notch-type OADM does not put a limit to extending the number of nodes in the ring. The FBG is made by writing a grating with UV light into the fibre core. By putting thermal or

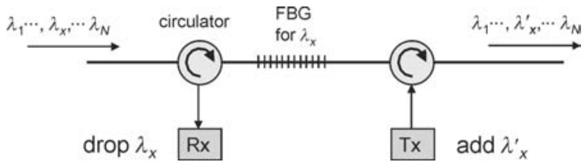

$\lambda_1 \cdots, \lambda_x, \cdots \lambda_N$
circulator
FBG
for λ_x
$\lambda_1 \cdots, \lambda'_x, \cdots \lambda_N$
drop λ_x
Rx
Tx
add λ'_x

Figure C1.2.20. Fixed-channel OADM using fibre Bragg grating (FBG). (Note: the right-hand circulator may also be replaced by a power combiner.)

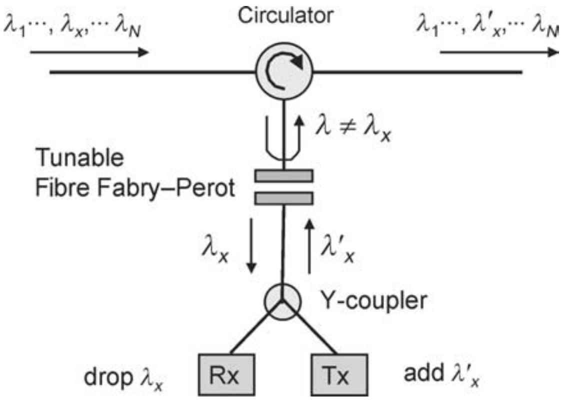

$\lambda_1\cdots, \lambda_x, \cdots \lambda_N$
Circulator
$\lambda_1\cdots, \lambda'_x, \cdots \lambda_N$
$\lambda \neq \lambda_x$
Tunable
Fibre Fabry–Perot
λ_x
λ'_x
Y-coupler
drop λ_x
Rx
Tx
add λ'_x

Figure C1.2.21. Tunable OADM using a fibre Fabry–Perot filter.

mechanical stress on the FBG, by means of a local heater or a piezo-electric stretcher, respectively, the device may be slightly tuned at low speed to other wavelength channels. Thus circuit-switched connections may be set up. Another more widely tunable OADM is shown in figure C1.2.21, which deploys a fibre Fabry–Perot (FFP) filter. The passband of the FFP can be tuned to any of the input wavelength channels, and thus the selected channel is dropped whereas the other channels are reflected without filtering and via the optical circulator passed to the output port. The locally added channel will pass the FFP and join the other channels via the circulator. The residual reflection (a few percent) at the FFP of the added channel causes near-end crosstalk at the local receiver; this may be counteracted by echo cancelling, as the locally added signal is known. An FFP is usually tuned with piezo-electric means; tuning speed is therefore limited, and this OADM is suited for setting up circuit-switched connections. Its notch-type characteristic implies that this OADM architecture does not limit the extension of the number of nodes in the ring.

When the wavelength channels are arranged in groups of closely spaced channels, adding and dropping of a specific group per node may be accomplished by a cascaded architecture of a fine-grain demultiplexer/multiplexer and a coarse grain OADM. Figure C1.2.22 shows an example, where a silica-based AWGR demultiplexer with narrow channel spacing and with a free spectral range of 500 GHz

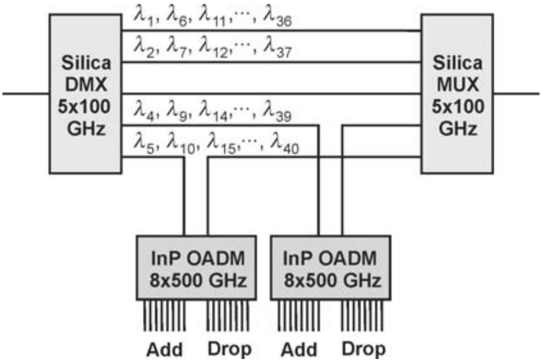

$\lambda_1, \lambda_6, \lambda_{11}, \cdots, \lambda_{36}$
$\lambda_2, \lambda_7, \lambda_{12}, \cdots, \lambda_{37}$
$\lambda_4, \lambda_9, \lambda_{14}, \cdots, \lambda_{39}$
$\lambda_5, \lambda_{10}, \lambda_{15}, \cdots, \lambda_{40}$
Silica DMX 5x100 GHz
Silica MUX 5x100 GHz
InP OADM 8x500 GHz
InP OADM 8x500 GHz
Add
Drop
Add
Drop

Figure C1.2.22. Two-stage OADM for handling wavelength groups.

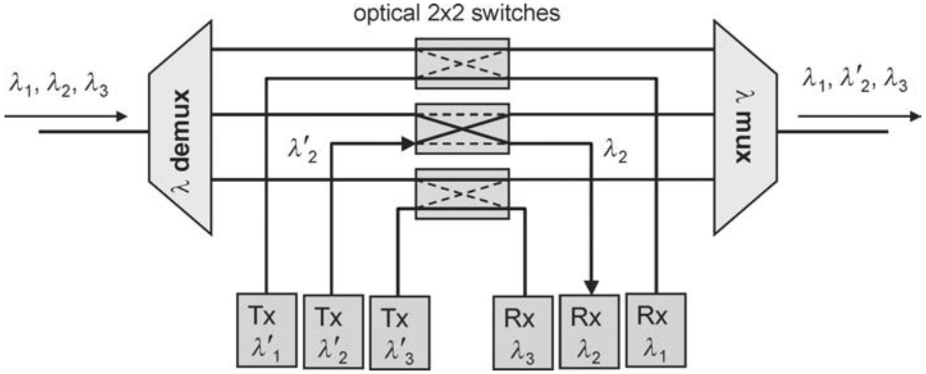


Figure C1.2.23. Fast wavelength-switchable OADM.

separates the 40 input wavelength channels spaced at 100 GHz into groups of eight wavelengths spaced at 500 GHz. Next, a compact OADM integrated in InP with a coarse channel spacing of 500 GHz separates the eight wavelength channels within a specific group.

For packet-switching applications, the OADM characteristics need to be fast tunable. Next to a wavelength demultiplexer stage and a wavelength multiplexer stage, the OADM architecture shown in figure C1.2.23 is equipped with fast (lithium-niobate, or semiconductor-based) optical switches, and can drop and/or add multiple wavelength channels. The wavelength passband channels of the demultiplexing and the multiplexing stages need to be and to stay carefully aligned. By putting a 2 × 2 optical switch in the cross state (as indicated for channel λ_2), the corresponding wavelength channel can be dropped and added. The pass-through wavelength channels are filtered, and thus this design limits the cascadability of nodes. An OADM architecture with similar functionality but requiring only a single wavelength-selective element is shown in figure C1.2.24 [8, 9]. The wavelength-selective AWGR performs both the demultiplexing and the multiplexing of the wavelength channels, and thus avoids wavelength misalignment issues. Some crosstalk may occur due to direct leak-through from the input port to the output port of the AWGR. By looping back the wavelength-demultiplexed paths via the add/drop switch matrix not to the front side but to the back side of the AWGR, and positioning the output port also at the front side, this crosstalk is strongly reduced; this fold-back architecture requires, however, a larger AWGR, with twice the number of ports at the back side.

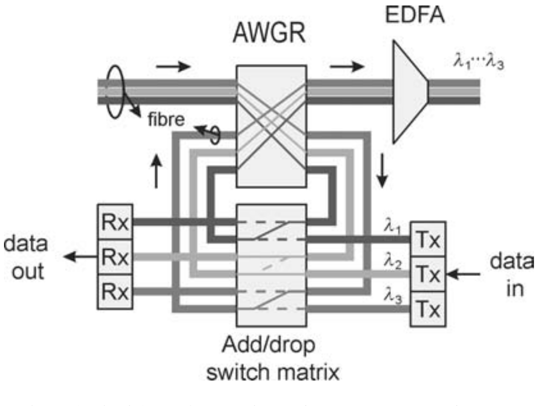


Figure C1.2.24. Fast wavelength-switchable OADM based on an arrayed waveguide grating router (AWGR).

C1.2.5 Access networks

Access networks carry a wide range of services to and from the residential end customers, ranging from voice-based services, audio-based ones, video-based ones, to Internet/data services. Also the capacities needed per service vary widely: from 64 kbit s^{-1} for traditional voice telephony to beyond 100 Mbit s^{-1} for high-speed Internet and data. The last link to connect to the end customer, the so-called first mile (or last mile, depending on the point of view), may be bridged with various types of transport media exploited by various network operators. Coaxial copper cable transports broadcast television and radio services, and increasingly also data services via cable modems. Twisted copper pair cables carry voice telephony, and data services via voice modems or high-speed ADSL and VDSL modems. Wireless systems bring mobile voice telephony via the GSM standard, and also data services via GPRS and UMTS. Optical fibre to the home/building is entering the market, but still has to surpass some cost barriers. It can offer the full set of integrated broadband services, from broadcast high bandwidth video services to gigabit Ethernet data services. A list of first-mile media with the bearer services, bitrates and reach is given in table C1.2.1. The need for more bandwidth in the access network is growing continuously, due to the increasing amount of bandwidth required by each customer, mainly fuelled by video-based services and high-speed Internet, the tailoring of services to individual customer needs, and the emergence of more competing operators due to liberalization. This spurs the introduction of optical fibre (mainly single mode fibre, being a future-proof solution with its virtually infinite bandwidth) into the access network. As the installation and equipment costs of fibre-to-the-home (FTTH) are still quite high in comparison to the traditional copper wired access lines, hybrid fibre access networks are the first step to introduce fibre. Fibre is used in the upper feeder part of the access network, where it runs from a local exchange (headend station) to a cabinet along the street (fibre-to-the-cabinet, FTTCab) or to

Table C1.2.1. First-mile network technologies.

Medium	Bearer service	Bitrate (down/up)	Reach (km)
Twisted pair	Analogue line	Rates up to 56 k/56 kbit s^{-1}	
Twisted pair	ISDN	144 k/144 k data incl. 64 k/64 kbit s^{-1} voice or data circuits	<6
Twisted pair	SDSL	768 k/768 kbit s^{-1}	<4
Twisted pair	ADSL	1.5 M to 6 M/64 k to 640kbit s^{-1}	<4–6
Twisted pair	VDSL	26 M to 52 M/13 M to 26 Mbit s^{-1}	<0.3–1
Coaxial cable	CDMA/OFDM + QAM/QPSK	<14 M/14 M (net 8.2 M) bit s^{-1} in 6 MHz slot	
Fibre (single mode)	ATM	150 M to 622 M/150 Mbit s^{-1} shared up to 1:32 (FSAN ATM-PON); up 1.24 G/620 Mbit s^{-1} (FSAN/ITU B-PON)	<20
Fibre (single mode)	Gigabit Ethernet	1 Gbit s^{-1} (1.25 Gbit s^{-1} 8B/10B coded)	<5
Fibre (multi mode)	Gigabit Ethernet	1 Gbit s^{-1} (1.25 Gbit s^{-1} 8B/10B coded)	<0.55
Wireless (mobile)	GSM	13 kbit s^{-1} (at carrier freq. 900 and 1800 MHz, freq. duplex)	<16
Wireless (mobile)	GPRS	115 kbit s^{-1}	
Wireless (mobile)	UMTS	144 k to 2 Mbit s^{-1} (at carrier freq. 2110–2200/1885–2025 MHz, freq. duplex)	
Wireless (fixed)	MMDS	6 Mbit s^{-1} (at carrier freq. >17 GHz)	
Wireless (fixed)	LMDS	45 Mbit s^{-1} (at carrier freq. >17 GHz)	

the basement of a building (such as an apartment building with many living units; fibre-to-the-building, FTTB), where the optical signals are converted back into electrical ones which are then brought via copper-based first mile links or wirelessly to the end customers. In the following of this section, the attention will be focussed on the optical fibre part of the access network.

Basically, three architectures may be deployed for the fibre access network:

(1) Point-to-point topology, where individual fibres run from the local exchange to each cabinet, home or building. Many fibres are needed, which entails high first installation costs, but also provides the ultimate capacity.

(2) Active star topology, where a single fibre carries all traffic to an active node close to the end users, from where individual fibres run to each cabinet/home/building. Only a single feeder fibre is needed, and a number of short branching fibres to the end users, which reduces costs; but the active node needs powering and maintenance.

(3) Passive star topology, in which the active node of the active star topology is replaced by a passive optical power splitter that feeds the individual short branching fibres to the end users. In addition to the reduced installation costs of a single fibre feeder link, the completely passive outside plant avoids the costs of powering and maintaining active equipment in the field. This topology has therefore become the most popular one for introduction of optical fibre into access networks, and is widely known as the passive optical network, PON.

C1.2.5.1 Multiple access PONs

The common fibre feeder part of the PON is shared by all the optical network units (ONUs) terminating the branching fibres. The traffic sent downstream from the optical line terminal (OLT) at the local exchange is simply broadcasted by means of the optical power splitter to every ONU. Sending traffic from the ONUs upstream to the local exchange, however, requires accurate multiple access techniques in order to multiplex collision-free the traffic streams generated by the ONUs onto the common feeder fibre. Four major categories of multiple access techniques for fibre access networks have been developed:

- Time division multiple access (TDMA).
- SubCarrier multiple access (SCMA).
- Wavelength division multiple access (WDMA).
- Optical code division multiple access (OCDMA).

In a TDMA system, as shown in figure C1.2.25, the upstream packets from the ONUs are time-interleaved at the power splitting point, which requires careful synchronization of the packet transmission instants at the ONUs. This synchronization is achieved by means of grants sent from the local exchange, which instruct the ONU when to send a packet. At the local exchange in the OLT, a burst mode receiver is needed which can synchronize quickly to packets coming from different ONUs, and which also can handle the different amplitude levels of the packets due to differences in the path loss experienced.

In an SCMA system, illustrated in figure C1.2.26, the various ONUs modulate their packet streams on different electrical carrier frequencies, which subsequently modulate the light intensity of the laser diode. The packet streams are thus put into different frequency bands, which are demultiplexed again at the local exchange. Each frequency band constitutes an independent communication channel from an

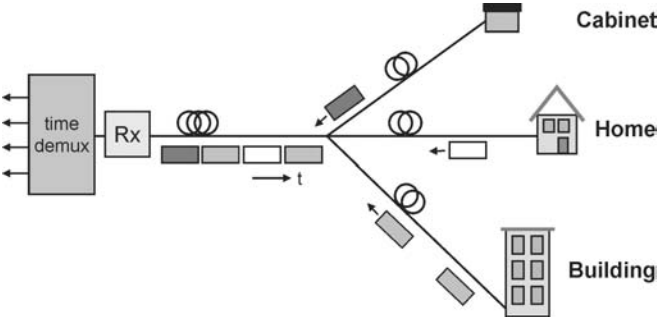


Figure C1.2.25. TDMA passive optical network.

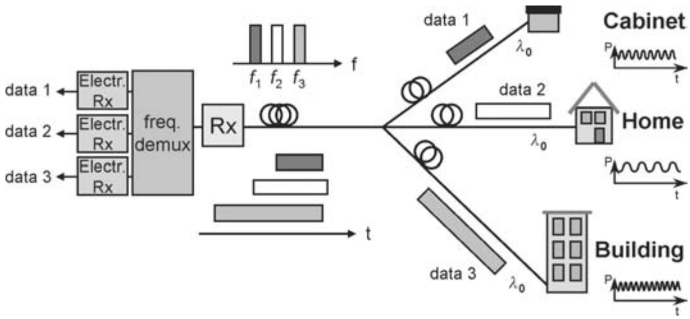


Figure C1.2.26. SCMA passive optical network.

ONU to the OLT in the local exchange and thus may carry a signal in a format different from that in an other channel (e.g. one channel may carry a high-speed digital data signal, and an other one an analogue video signal). No time synchronization of the channels is needed. The laser diodes at the ONUs may have nominally the same wavelength. When the wavelengths of the lasers are very close to each other, the frequency difference between them may result in beat noise products due to optical beating at the photodetector in the receiver. These noise products may interfere with the packet data spectrum. The wavelengths of the laser diodes have to be adjusted slightly different (e.g. by thermal tuning) in order to avoid this optical beat noise interference.

In a WDMA system (see figure C1.2.27), each ONU uses a different wavelength channel to send its packets to the OLT in the local exchange. These wavelength channels constitute independent communication channels and thus may carry different signal formats; also no time synchronization is needed. The same wavelength channel may be used for upstream communication as for downstream. The isolation requirements of the wavelength demultiplexer may be high to sufficiently suppress crosstalk, e.g. when high-speed digital data and analogue video are carried on two different wavelength channels. The channel routing by the wavelength multiplexer at the network splitting point prohibits broadcasting some channels to all ONUs, as needed for instance for CATV signal distribution. Every ONU needs a wavelength-specific laser diode, which increases costs, and complicates maintenance and stock inventory issues. An alternative is to use a light source with a broad spectrum at the ONU (e.g. a superluminescent LED), of which the in-field multiplexer cuts out the appropriate part of the spectrum. This 'spectral slicing' approach reduces the inventory problems, but also yields a reduced effective optical power available from the ONU and thus limits the reach of the system. Another alternative is to use a reflective modulator at the ONU, which modulates the upstream data on a continuous light

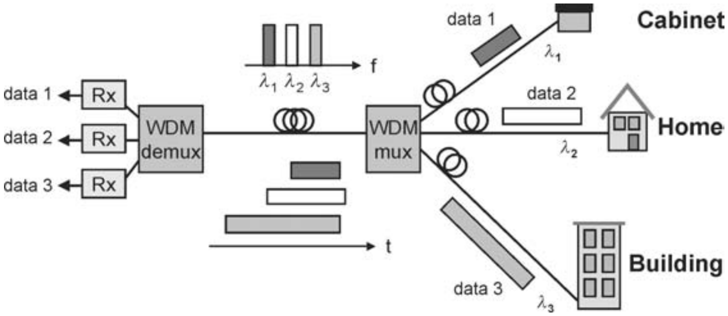


Figure C1.2.27. WDMA passive optical network.

channel emitted at the appropriate wavelength by the OLT and returns it to the OLT [10]. Thus no light source is needed at the ONU, which eases maintenance; but again the power budget is limited.

In an OCDMA system, each ONU uses a different signature sequence of optical pulses, and this sequence is on-off modulated with the data to be transmitted. The duration of the sequence needs to be at least equal to that of a data bit, and thus a very high-speed signature sequence is needed to transmit moderate-speed data which limits the reach of the system due to the increased impact of dispersion and the decreasing power budget at high line rates. In the OLT at the local exchange, the received signals are correlated with the known signature sequences, in order to demultiplex the data coming from the different ONUs. As the signature codes may not be perfectly orthogonal, some crosstalk may occur.

TDMA systems have received the most attention for broadband access networks, as they are most suited for high-speed data transmission at relatively moderate complexity. Two types of TDMA passive optical networks have been addressed extensively in standardization bodies: the ATM PON (APON) carrying native ATM cells in the G.983 standard series of ITU-T SG15, and the Ethernet PON (EPON) carrying gigabit Ethernet packets in IEEE 802.3.

C1.2.5.2 ATM PON

The full service access network (FSAN) group, a committee of presently 21 major telecommunication operators around the world, since 1995 is promoting the ATM PON for broadband access networks.

As laid down in the G.983.1 Recommendation of ITU-T [11], an ATM PON may have a downstream bitrate of 155 or 622 $\mathrm{Mbit\,s^{-1}}$, and an upstream one of 155 $\mathrm{Mbit\,s^{-1}}$. The maximum optical splitting ratio is 32 (may grow to 64), and the maximum fibre length between the OLT in the local exchange and an ONU is 20 km. The range in which this length is allowed to vary is from 0 to 20 km. Standard single mode fibre (G.652) is foreseen. Coarse wavelength multiplexing is used for separating the bi-directional traffic: the downstream traffic is positioned in the 1.5 μm wavelength band, and the upstream traffic in the 1.3 μm band (using cheap Fabry–Perot laser diodes in the ONUs).

In the downstream direction of a 155 $\mathrm{Mbit\,s^{-1}}$ down/155 $\mathrm{Mbit\,s^{-1}}$ up system, 54 ATM cells of 53 bytes each are fitted together with two physical layer operation, administration, and maintenance (PLOAM) cells of 53 bytes in a frame [11]. The PLOAM cells contain each 53 upstream grants. A grant permits an ONU to send an ATM cell. By sending these grants, the OLT controls at each ONU the transmission of the upstream packets, and can therefore assign dynamically a portion of the upstream bandwidth to each ONU. In a 622 $\mathrm{Mbit\,s^{-1}}$ down/155 $\mathrm{Mbit\,s^{-1}}$ up system, a frame contains four times as many cells (i.e. 216 ATM cells and eight PLOAM cells). The downstream frame is broadcasted to all ONUs. An ONU only extracts those cells that are addressed to it.

In the upstream frame, both for the 155 Mbit s^{-1} down/155 Mbit s^{-1} up system and for the 622 Mbit s^{-1} down/155 Mbit s^{-1} up system, 53 ATM cells are fitted of 53 bytes each plus an overhead of 3 bytes per cell. This overhead is used as guard time, as a delimiter and as preamble for supporting the burst mode receiver process in the local exchange.

The power budgets needed to bridge the fibre losses and the splitter losses are denoted by three classes of optical path losses: class A 5–20 dB, class B 10–25 dB, and class C 15–30 dB. At the ONU, a launched optical power of −4 to +2 dBm is specified for class B, and −2 to +4 dBm for class C [12]. The ONU receiver sensitivity at 155 Mbit s^{-1} should be better than −30 dBm for class B, and −33 dBm for class C.

The ONUs are usually positioned at different distances from the local exchange. Therefore, the upstream transmission of the packets from each ONU should be carefully timed, in such a way that the packets do not collide at the network splitter [11, 14]. The OLT has to measure the distance to each ONU for this, and then instructs the ONU to insert an equalizing transmission delay such that all distances from the ONUs to the OLT are virtually equal to the longest allowable distance (i.e. 20 km); see figure C1.2.28. To measure the distance to each ONU, the OLT emits a ranging grant to each ONU, and on receipt the ONU returns a ranging cell to the OLT. In this distance ranging process, the OLT can deduce the distance to each ONU from the round trip delay.

Each ONU sends an upstream cell upon the receipt of a grant. Because the path losses from each ONU to the OLT may be different, the power of the cells received by the OLT may vary considerably from cell to cell. The burst mode receiver at the OLT should therefore have a wide dynamic range, and should be able to set its decision threshold quickly to the appropriate level to discriminate the logical ones from the zeros. Also the power of the ONU transmitter can be varied over a certain range to limit the requirements on the receiver dynamic range. In this amplitude ranging process, the overhead to each ATM cell is used for supporting the fast decision threshold setting at the OLT burst mode receiver and the power adaptation at the ONU burst mode transmitter.

Four types of network protection have been described in Recommendation G.983.1 [12], as shown in figure C1.2.29. Type A protection involves protection of the feeder fibre only by a spare fibre over which the traffic can be rerouted by means of optical switches. After detection of a failure in the primary fibre and switch-over to the spare fibre, also re-ranging has to be done by the PON transmission convergence (TC) layer. Thus only limited protection of the system is realized. Mechanical optical

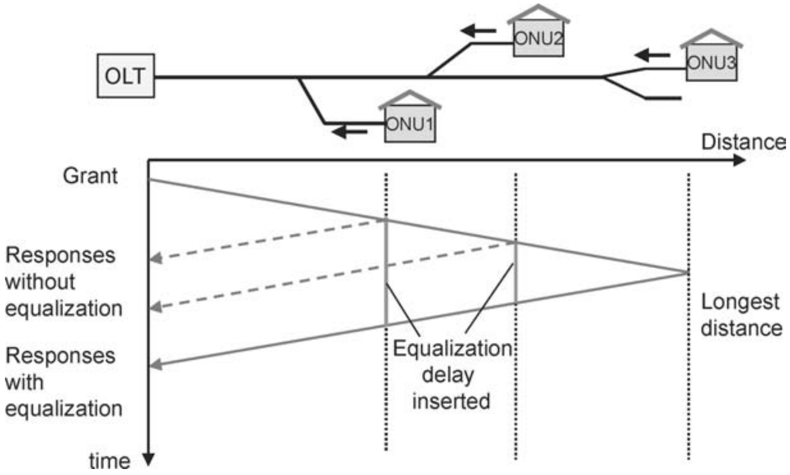
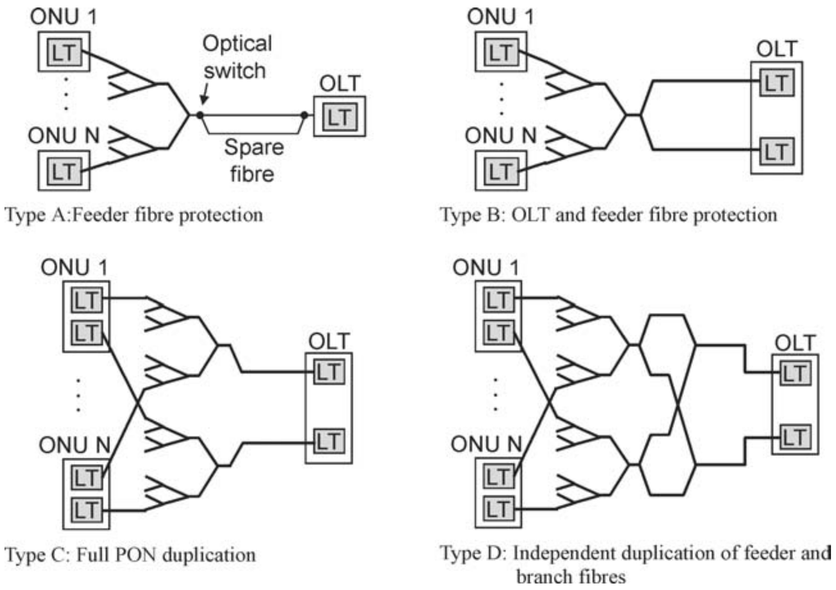


Figure C1.2.28. Time ranging in a TDMA PON.



Type A:Feeder fibre protection

Type B: OLT and feeder fibre protection

Type C: Full PON duplication

Type D: Independent duplication of feeder and branch fibres

Figure C1.2.29. PON protection schemes.

switches are used up to now; when optical switching becomes cheaper, this protection scheme may become more attractive. Type B protection features duplication of both the feeder fibre and the OLT. The secondary OLT is on cold standby, and is activated when the primary one fails. Due to the high sharing factor of the duplicated resources by the ONUs, this approach offers an economical yet limited protection. Type C protection implies full duplication of the PON, and all equipment is normally working which allows fast switch-over (within 50 ms) from the primary equipment to the secondary one. The branch fibres as well as the ONUs are protected; also a mix of protected and unprotected ONUs can be handled. Type D protection features independent duplication of the feeder fibres and the branch fibres. It cannot offer fast restoration. It is less attractive than C, as it requires more components but not a better functionality. In summary, types B and C are the most attractive schemes for a new recommendation.

To further increase the speeds laid down in Recommendation G.983.1, research is done into 622, 1244 and 2488 Mbit s^{-1} line rates, both for upstream and downstream. A key technical issue is the development of faster burst-mode circuitry to adequately retrieve the timing and set the decision threshold level, which becomes increasingly more difficult at higher line rates. Operation of 622 Mbit s^{-1} burst-mode circuitry has been achieved recently [12]. In January 2003, ITU has set standards for gigabit-capable PONs (G-PONs). These ITU-T Recommendations G.984.1 and G.984.2 cover downstream speeds of 1.25 and 2.5 Gbit s^{-1}, and upstream speeds of 155 and 622 Mbit s^{-1}, and of 1.25 and 2.5 Gbit s^{-1}.

The G.983.1 ATM PON was initially mainly designed for high-speed data communication. However, in the residential access networks there is also a clear demand for economical delivery of CATV services, for which subcarrier multiplexing techniques are quite appropriate. In the enhanced Recommendation G.983.3 [12], room has been allocated in the optical spectrum to host video services or additional digital services next to the ATM PON services. As shown in figure C1.2.30, the APON upstream services remain in the 1260–1360 nm band (as in G.983.1), but the band for downstream services is narrowed to 1480–1500 nm (1480–1580 nm in G.983.1). Next to those, an enhancement band for densely wavelength multiplexed bi-directional digital services (such as private wavelength services) is

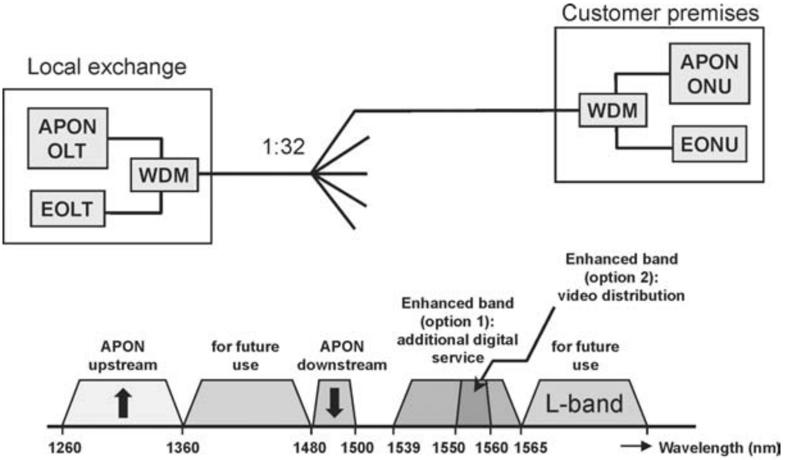


Figure C1.2.30. WDM enhancement G.983.3.

foreseen, or an enhancement band for an overlay of video delivery services. The latter is used in downstream direction only, and coincides with the C-band as thus economical erbium-doped fibre amplifiers (EDFAs) can be deployed for the power boosting required. When positioning an overlay of CATV distributive services in the C-band, stringent crosstalk requirements have to be put on the wavelength multiplexers and demultiplexers, to prevent noticeable interference of the CATV signals into the digital ATM signals, and vice versa [13].

In order to further improve the economics of ATM PON systems, an extended PON system with an increase of the network splitting factor to 128 and even 256 has been developed, while still maintaining a passive outside plant and compatibility with G.983.1 compliant ONUs [14]. This extended split is achieved by a larger optical power budget. In the downstream direction, at the OLT a high power laser diode or an EDFA is used to boost the power. In the upstream direction, the sensitivity of the burst-mode receiver is improved by applying an avalanche photo diode (APD). Also eight single-mode feeder fibres (each feeding a 1:16 or 1:32 power splitter in the field) are at the OLT coupled to a multimode fibre yielding a low-loss coupling to the receiver.

Even further extensions of the split factor and of the reach of an ATM PON have been realized in the SuperPON system [15]. An extension to a splitting factor of 1:2048 has been achieved; this needs, however, active equipment in the field. In the downstream direction exploiting the 1530–1560 nm wavelength window, EDFAs are used for overcoming the large path losses. In the upstream direction, gated semiconductor optical amplifiers (SOAs) are deployed. Each SOA gate is opened when upstream packets arrive, and is shut otherwise in order to avoid funneling of the amplified spontaneous emission noise towards the OLT. This SuperPON approach is not compliant with present standards, and may be economically feasible only in the long term [14].

C1.2.5.3 Ethernet PON

With the rapid penetration of Ethernet-based services, Ethernet PON (EPON) techniques are receiving increasing attention, and are promoted by the IEEE 802.3 Ethernet in the first mile (EFM) group. The major difference with ATM PONs is that an EPON carries variable-length packets up to 1518 bytes in length, whereas an ATM PON carries fixed-length 53 bytes cells. This ability yields a higher efficiency for handling IP traffic. The packets are transported at the gigabit Ethernet 1.25 $\mathrm{Gbit\,s^{-1}}$ speed using the IEEE 802.3 Ethernet protocol. However, ATM offers built-in quality of service for all traffic classes,

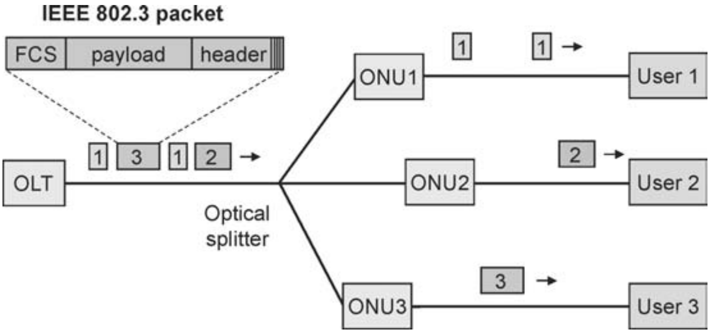

IEEE 802.3 packet
FCS
payload
header
OLT
1
3
1
2
Optical splitter
ONU1
ONU2
ONU3
User 1
User 2
User 3

Figure C1.2.31. Downstream traffic in an EPON.

whereas Ethernet does not. EPON thus cannot support voice services with quality of service as provided in the traditional public switched telephone network (PSTN), and also the support of real-time services still has issues due to latency and packet jitter.

The EPON features full-duplex transmission similarly as the ATM PON, with downstream traffic at 1490 or 1510 nm, and upstream traffic at around 1310 nm. As shown in figure C1.2.31, standard IEEE 802.3 Ethernet packets are broadcasted downstream by the OLT to all the ONUs. Each ONU inspects the headers, and extracts the packets that are addressed to it. Several variable-length packets are put into a fixed-length frame of 2 ms duration, and each frame begins with a 1-byte synchronization marker. In the upstream direction, also 2 ms frames are used. A frame contains time slots that each are assigned to one of the ONUs (see figure C1.2.32). Each ONU puts one or more of its upstream variable-length IEEE 802.3 packet into a time slot; if it has no packets to send, the time slot may be filled with an idle signal. No packet fragmentation takes place. The time slot overhead consists of a guard band, and indicators for timing and signal power. The OLT thus allows only one ONU to send at a time, and no collisions occur. The time slot size is 125 or 250 μs.

C1.2.5.4 Hybrid fibre coax networks

CATV networks usually are laid out over large geographical areas, and are mainly designed for downstream broadcasting of analogue TV channels that are frequency-division multiplexed in a carrier frequency grid extending up to 1 GHz. As shown in figure C1.2.33, in a hybrid fibre coax (HFC) system

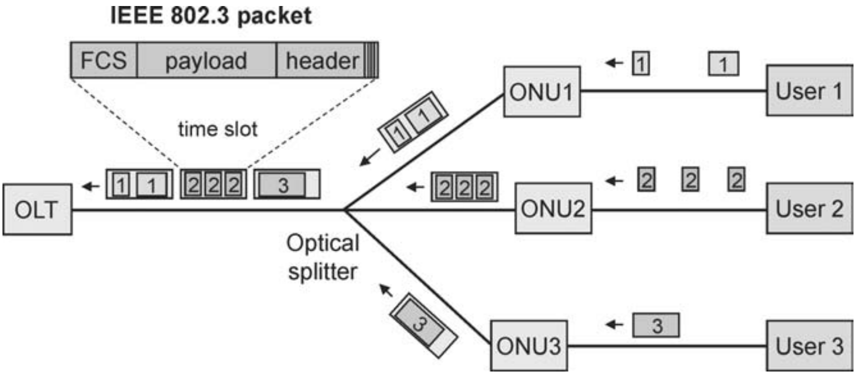

IEEE 802.3 packet
FCS
payload
header
time slot
OLT
1
1
2
2
2
3
Optical splitter
ONU1
ONU2
ONU3
User 1
User 2
User 3

Figure C1.2.32. Upstream traffic in an EPON.

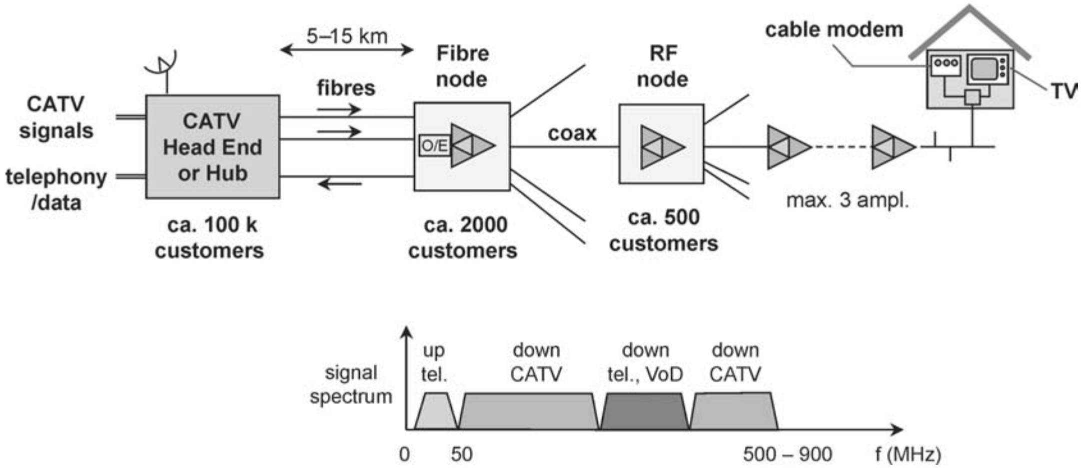

5–15 km
fibres
CATV signals
telephony /data
CATV Head End or Hub
ca. 100 k customers
Fibre node
O/E
ca. 2000 customers
coax
RF node
ca. 500 customers
cable modem
TV
max. 3 ampl.
signal spectrum
up tel.
down CATV
down tel., VoD
down CATV
0
50
500 – 900
f (MHz)

Figure C1.2.33. Hybrid fibre coax network.

a CATV headend station is collecting the CATV signals, remodulating them into a specific frequency grid, and sending them via single-mode fibres to fibre nodes. Each fibre node converts the composite optical signal into an electrical one, which is carried via a coaxial cable network including several RF amplifiers to the residential homes. A single headend may thus serve hundred thousands of customers, and a fibre node some thousands of customers. In particular during transmission in the coaxial cable network, the signal quality deteriorates due to the addition of noise from the electrical amplifiers and intermodulation products from nonlinearities in the system. On the fibre part of the network, the signals are carried with subcarrier multiplexing; see figure C1.2.34. The TV channels each are amplitude-modulated on a separate frequency, and after summing all these modulated signals, a highly linear high

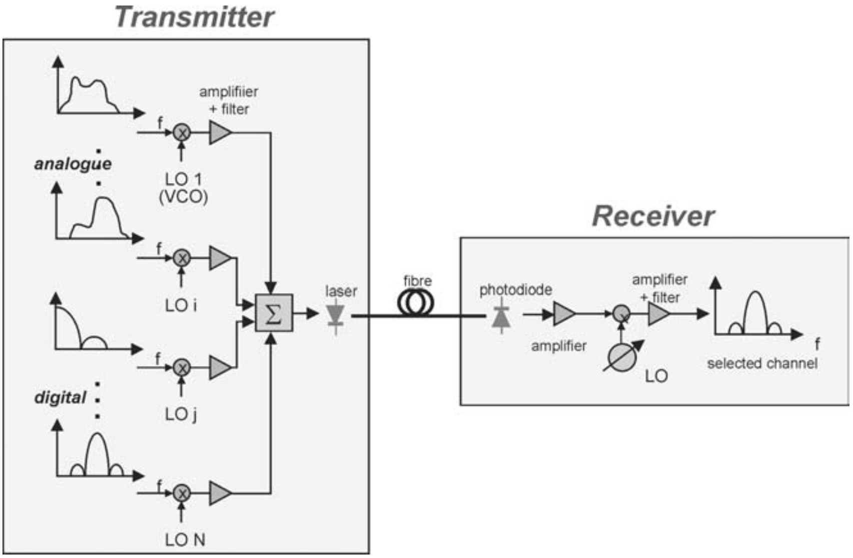

Transmitter
amplifiier + filter
f
LO 1 (VCO)
analogue
LO i
LO j
digital
LO N
Σ
laser
fibre
Receiver
photodiode
amplifier
amplifier + filter
LO
f
selected channel

Figure C1.2.34. Subcarrier multiplexing.

power laser diode (or laser diode followed by a linearized external modulator) generates an optical signal which is intensity-modulated with the composite CATV signal. At the receiver site, the optical signal is converted into the electrical CATV signal by means of a highly linear PIN photodiode, and subsequently the signal can be passed to the coaxial cable network or to a selective receiver. When using a laser diode with low relative intensity noise and high linearity (or a carefully linearized external modulator), the CATV signal can be transported with very little loss of quality. If a 1.5 μm wavelength laser diode is used, EDFAs may boost the power at the headend and compensate for the splitting losses; thus very extensive networks feeding thousands of ONUs can be realized. In this wavelength region, however, with direct laser modulation second order intermodulation products may arise due to laser chirp in combination with fibre chromatic dispersion; with an external modulator, however, the chirp is small enough to avoid these intermodulation products.

The CATV signal quality that can be maintained in HFC networks is very high due to the fibre's low losses and high bandwidth in comparison with coaxial cable. Therefore, in HFC networks fibre is gradually brought deeper into the network, and fibre nodes have to serve fewer customers through a coaxial cable network of limited size (i.e. mini fibre nodes, each serving in the order of 40 customers).

At present, HFC networks are not only carrying CATV and FM radio broadcast services, but cable operators are also exploiting them for voice telephony and data transport using cable modems. For the upstream traffic involved with these interactive services, parts of the spectrum unused for CATV and FM radio broadcast can be used. In Europe, typically the 5–65 MHz band is used for this; in the US, the 5–42 MHz range. For downstream data, e.g. the 300–450 MHz range is used, taking into account that Internet traffic is usually highly asymmetric (much more downloading traffic than uploading). Downstream per 8 MHz CATV channel, 30–50 Mbit s^{-1} data can be accommodated deploying 64 or even 256 quadrature amplitude modulation (QAM). Upstream due to ingress noise less complicated modulation schemes are to be used; DQPSK offers about 3 Mbit s^{-1} per channel.

C1.2.5.5 Dense wavelength multiplexing in access networks

In general, access networks have to meet a fast growth in capacity demand, due to several causes: customers are asking for second and more telephone lines; Internet data traffic is booming with higher data rates, more users and longer sessions on-line (even always on); an increasing amount of video-based services; fast growth in number of mobile phone users and session frequency; new operators entering asking to rent capacity on existing access networks; etc. This hunger for more capacity and the strive for convergence of services on a single network can most adequately be met by bringing fibre ever closer to the end users, from where only a short copper cable based (or wireless) link has to be bridged to the customer. Ultimately, when installation and equipment costs have come down sufficiently, the most powerful network is achieved when fibre runs all the way to the customer's home (fibre to the home, FTTH).

The upgradation of installed fibre plant to higher capacities while protecting the investments made is efficiently done by introducing wavelength multiplexing techniques. Wavelength channels may be allocated to specific sets of services (for service unbundling), and/or to separate service operators (leasing of network capacity).

Dynamic capacity allocation by flexible wavelength assignment

To cope with variation in service demand by the users and the sometimes quickly changing operator conditions, it is more efficient to flexibly allocate the augmented available network capacity across the access network. Dynamic wavelength routing techniques can be used for this, thus making more efficient use of the network's resources and generating more revenues. Figure C1.2.35 illustrates the principle: from the OLT in the headend station of the network, multiple wavelength channels are fed to the ONUs

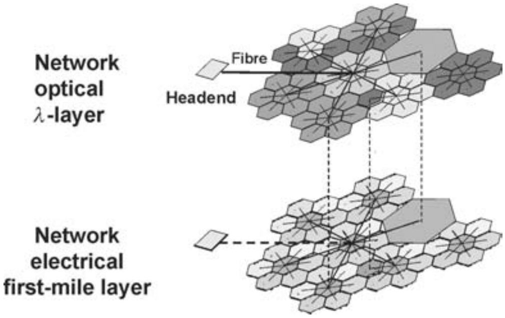
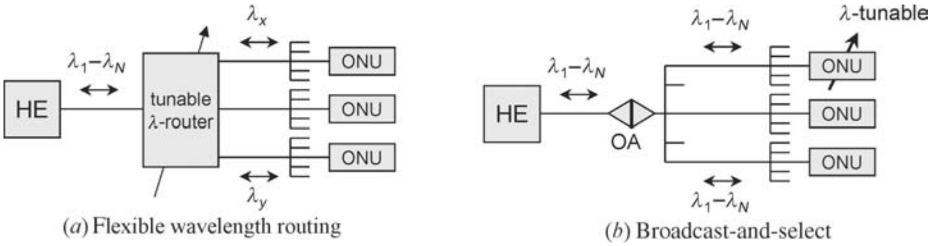


Figure C1.2.35. Dynamic wavelength routing in hybrid access networks.

via a tree-and-branch PON. By wavelength-selective routing in the PON, or wavelength selection at the ONU, wavelength channels can be assigned to a number of specific ONUs. Thus capacity can be specifically shared between these ONUs. The ONUs subsequently transfer this capacity shares to their first-mile electrical network connecting the end users. The mapping of the network capacity resources to the first-mile networks can thus be changed by changing the wavelength channel assignment. Basically two approaches can be followed, as illustrated in figure C1.2.36: a wavelength router in the field, or wavelength selection at the ONUs. As shown in figure C1.2.36(*a*), a tunable wavelength router directs the wavelength channels to specific output ports, and this routing can be dynamically adjusted by external control signals from the headend. In order to support in addition the delivery of broadcast services to all ONUs, extra provisions have to be made for enabling broadcast wavelength channel(s) to bypass the router. As the wavelength channels are routed to only those ONUs whose customers require the associated services, no optical power is wasted. As shown in figure C1.2.36(*b*), another approach is to broadcast all wavelength channels to every ONU, and subsequently tune the ONU to the wavelength channel wanted. Clearly the power of the other wavelength channels is wasted by the ONU, and losses at the broadcasting power splitter are significant. An optical amplifier is usually needed to make up for these losses; the amplifier needs to operate bi-directionally to handle downstream as well as upstream traffic. No specific provisions in the network are needed for supporting broadcast services.

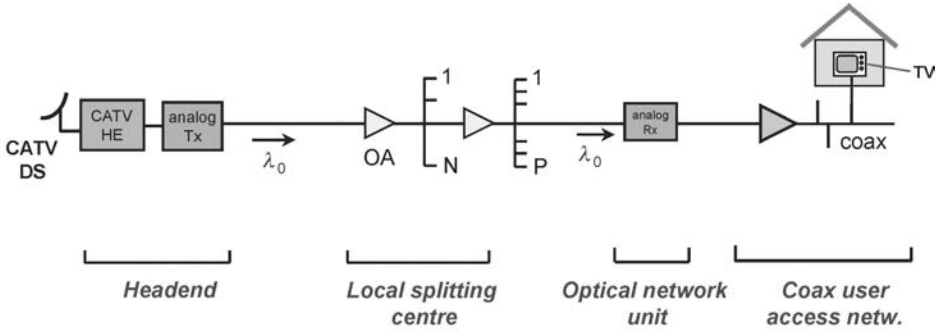
Figure C1.2.37 presents a multi-wavelength overlay of a number of ATM PON networks on a HFC network, following the *wavelength channel selection* approach [16]. Figure C1.2.37(*a*) shows a fibre-coax network for distribution of CATV services, operating at a wavelength λ_0 in the 1550–1560 nm window where EDFAs offer their best output power performance. Thus, using several EDFAs in cascade, an extensive optical network splitting factor can be realized and a large number of customers can be served. For example, with two optical amplifier stages and typical splitting factors of $N = 4$ and $P = 16$, and a mini-fibre node serving 40 users via its coaxial network, a total of 2560 users is served from a single headend fibre. For interactive services, the upstream frequency band in a standard HFC network



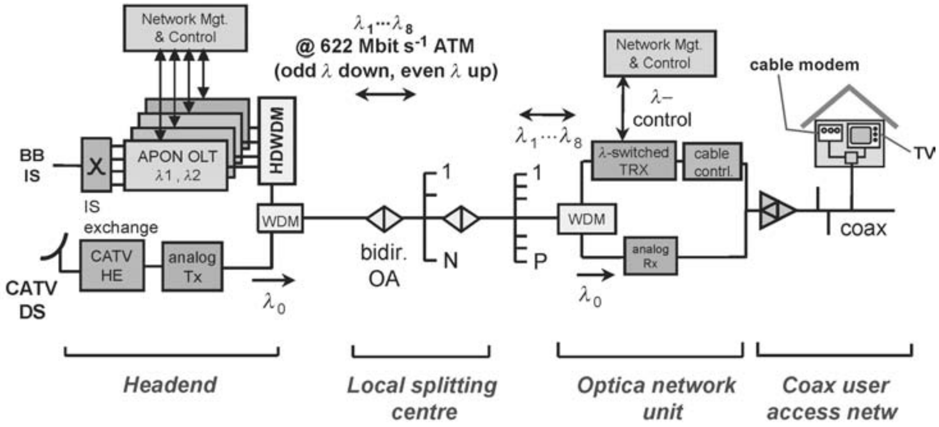
(*a*) Flexible wavelength routing

(*b*) Broadcast-and-select

Figure C1.2.36. Dynamically allocating wavelength channels to ONUs.


CATV DS
CATV HE
analog Tx
λ0
OA
1
N
1
P
λ0
analog Rx
TV
coax
Headend
Local splitting centre
Optical network unit
Coax user access netw.

(*a*) fibre-coax network for distribution of CATV services


Network Mgt. & Control
λ1···λ8
@ 622 Mbit s-1 ATM
(odd λ down, even λ up)
HDWDM
BB IS
X
APON OLT
λ1, λ2
IS exchange
CATV HE
analog Tx
CATV DS
WDM
λ0
bidir. OA
1
N
1
P
λ1···λ8
Network Mgt. & Control
λ–control
cable modem
λ-switched TRX
cable contrl.
TV
WDM
analog Rx
λ0
coax
Headend
Local splitting centre
Optica network unit
Coax user access netw

(*b*) Upgrading of the fibre -coax network with multi-wavelength APON system for delivery of broadband interactive services

Figure C1.2.37. Flexible capacity assignment in a multi-wavelength fibre-coax network by wavelength selection at the ONUs.

(with a width of some 40–60 MHz) has to be shared among these users, thus allowing only limited bitrates per user for narrowband services such as voice telephony.

An upgrade of the system in order to provide broadband interactive services can be realized by overlaying the HFC network with a number of wavelength-multiplexed APON systems, as developed in the ACTS TOBASCO project [16] and shown in figure C1.2.37(*b*). Four APON OLTs at the headend site are providing each bi-directional 622 Mbit s^{-1} ATM signals on a specific downstream and upstream wavelength. These eight wavelengths are positioned in the 1535–1541 nm window, where the up- and downstream wavelength channels are interleaved with 100 GHz spacing. The APON wavelengths are combined by a high-density wavelength division multiplexer (HDWDM), and subsequently multiplexed with the CATV signal by means of a simple coarse wavelength multiplexer (due to the wide spacing between the band of APON wavelengths and the CATV wavelength band). The system upgrade implies also replacement of the uni-directional optical EDFAs by bi-directional ones which feature low noise high-power operation for the downstream CATV signal, and for the bi-directional ATM signals a wavelength-flattened gain curve plus a nonsaturated behaviour (to suppress crosstalk in burst-mode). At the ONU site, first the CATV signal is separated from the APON signals by means of a coarse wavelength multiplexer, and is subsequently converted to an electrical CATV signal by a highly linear receiver and distributed to the users via the coaxial network. The APON signals are fed to a wavelength-switched transceiver, of which the receiver can be switched to any of the four downstream wavelength channels, and the transmitter to any of the four upstream ones. The wavelength-switched transceiver may be implemented by an array of wavelength-specific transmitters and receivers, which can be individually switched on and off; this configuration allows to set up a new wavelength channel before breaking down the old one ('make-before-break'). Alternatively, it may use wavelength-tunable transmitters and receivers, which can in principle address any wavelength in a certain range; this eases further upgrading of the system by introducing more wavelength channels, but also implies a 'break-before-make' channel switching. The network management and control system commands to which downstream and to which upstream wavelength channel each ONU transceiver is switched. By issuing these commands from the headend station, the network operator actually controls the virtual topology of the network, and thus is able to allocate the network's capacity resources in response to the traffic demands at the various ONU sites. The network management command signals are transported via an out-of-band wavelength channel in the 1.3 μm wavelength window. The APON signal channel selected by the ONU is converted into a bi-directional electrical broadband data signal by the transceiver, which is by a cable modem controller put in an appropriate frequency band for multiplexing with the electrical CATV signal. The upstream data signal is usually put below the lowest frequency CATV signal (so below 40–50 MHz), and the downstream signal in empty frequency bands between the CATV broadcast channels. The signals are carried by the coaxial network (in which only the electrical amplifiers need to be adapted to handle the broadband data signals) to the customer homes, where the CATV signal is separated from the bi-directional data signals; the latter signals are processed by a cable modem, which interacts with the cable modem controller at the ONU site.

By remotely changing the wavelength selection at the ONUs, the network operator can adjust the system's capacity allocation in order to meet the local traffic demands at the ONU sites. As illustrated in figure C1.2.39, the ONUs are allocated to the four upstream (and downstream) wavelength channels, which each have a maximum capacity of 622 Mbit s^{-1} for ATM data. As soon as the traffic to be sent upstream by an ONU grows and does not fit anymore within its wavelength channel, the network management system can command the ONU to be allocated to an other wavelength channel, in which still sufficient free capacity is available. Obviously, this dynamic wavelength re-allocation process reduces the system's blocking probability, i.e. it allows the system to handle more traffic without blocking and thus can increase the revenues of the operator.

Figure C1.2.38 presents the *dynamic wavelength channel routing* approach in a fibre-wireless network to allocate flexibly the capacity of a number of ATM PON systems among ONUs in a single fibre split network infrastructure [17]. The ONUs are each feeding a radio access point (RAP) of e.g. a wireless LAN, which wirelessly connects to a variable number of users with mobile terminals. These users move across the geographical area served by the network (e.g. a business park), and they may want to set up a broadband wireless connection to their laptop at any time anywhere in this area. When many users are within a wireless cell served by a certain RAP, this cell may have to handle much more traffic than the other cells; it has become a 'hot spot' which has to be equipped with additional capacity. The corresponding RAP may switch on more microwave carriers to provide this additional capacity over the air, and also has to claim more capacity from the ONU. This local extra capacity can be provided by re-allocation of the wavelength channels over the ONUs, which is done by a flexible wavelength router positioned in the field. Similar to the architecture of the wavelength-reconfigurable fibre-coax network in figure C1.2.37(*b*), the architecture in figure C1.2.38 developed in the ACTS PRISMA project has four 622 Mbit s^{-1} bi-directional APON OLTs with a specific downstream wavelength and an upstream one each. The four downstream wavelengths are located in the 1538–1541 nm range, with 100 GHz spacing, and the four upstream ones in the 1547–1550 nm range with the same spacing. The flexible wavelength router directs the downstream wavelength channels each to one or more of its output ports, and thus via

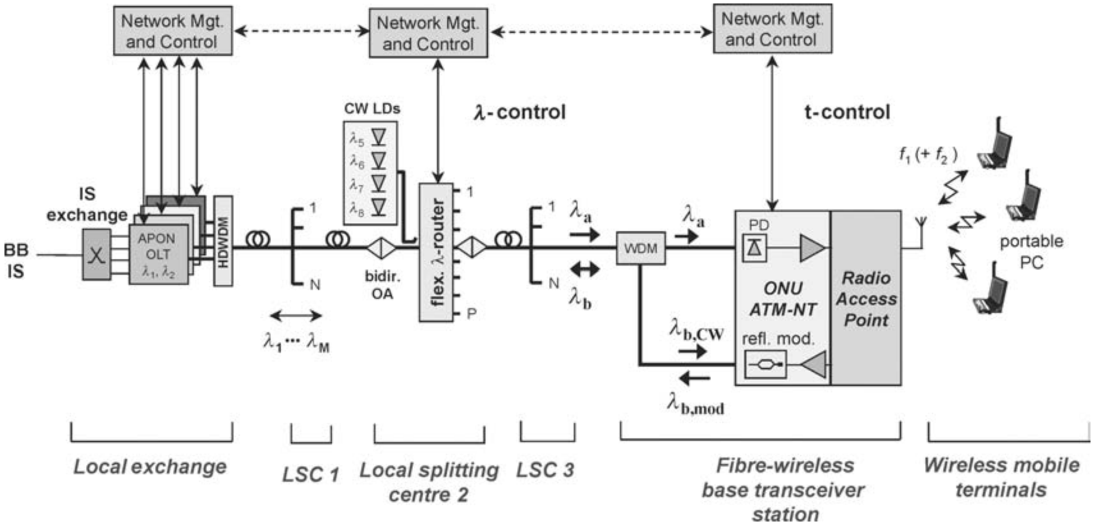

Network Mgt. and Control
Network Mgt. and Control
Network Mgt. and Control
λ-control
t-control
CW LDs
λ5
λ6
λ7
λ8
IS exchange
BB IS
APON OLT λ1, λ2
HDWDM
1
N
bidir. OA
flex. λ-router
1
P
λ1 ··· λM
1
N
λa
λb
WDM
λa
λb,CW
λb,mod
PD
ONU ATM-NT
refl. mod.
Radio Access Point
f1 (+ f2)
portable PC
Local exchange
LSC 1
Local splitting centre 2
LSC 3
Fibre-wireless base transceiver station
Wireless mobile terminals

Figure C1.2.38. Flexible capacity assignment in a multi-wavelength fibre-wireless network by wavelength routing in the field.

a split network to a subset of ONUs. The RAPs could operate with up to five microwave carriers in the 5 GHz region, each carrying up to 20 Mbit s^{-1} ATM wireless LAN data in OFDM format. At the flexible router (or at the local exchange) a number of continuous-wave emitting laser diodes are located, which provide unmodulated light power at the upstream wavelengths. The flexible router can select one of these upstream wavelengths, and direct it to the ONUs that can modulate the signal with the upstream data and return it by means of a reflective modulator via the router to the local exchange. Thus no wavelength-specific source is needed at the ONU, the downstream light sources are shared by a number of ONUs, and all ONUs are identical, which reduces the system costs and the inventory issues. The flexible wavelength router can be implemented with a wavelength demultiplexer separating the wavelength channels, followed by power splitters, optical switches and power couplers in order to guide the channels to the selected output port(s). Depending on the granularity of the wavelength allocation process, the flexible router may be positioned at desired different splitting levels in the network.

Using a similar strategy to assign wavelength channels to the ONUs as shown in figure C1.2.39, a statistical performance analysis has been performed of the blocking probability of the system. It was assumed that the total network served 343 cells, of which 49 were ‘hot spots’, i.e. generated a traffic load two times as large as a regular cell. It was also assumed that the system deployed seven wavelength channels, and that the calls arrived according to a Poisson process where the call duration and length were uniformly distributed. Figure C1.2.40 shows how the system blocking probability depends on the offered load (normalized on the total available capacity, which is 7 times 622 Mbit s^{-1}), using various system architecture options. When wavelength re-allocation would not be possible (i.e. static WDM) and all the 49 hot spots were positioned at cells served by ONUs assigned to the same wavelength channel, the blocking probability is obviously the worst case. On the other hand, in the static WDM case when the 49 hot spots were evenly spread over the seven wavelength channels, the blocking probability is much lower (i.e. best case). Unfortunately, a network operator cannot know beforehand where the hot spots will be positioned, so in this static WDM situation the system blocking probability will be anywhere between the best case and the worst case, and no guarantee for a certain blocking performance can be given. When, however, dynamic re-allocation of the wavelength channels is possible, the system

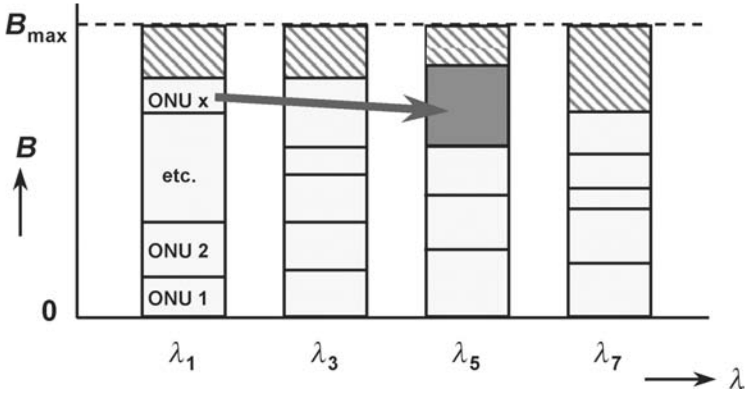


Figure C1.2.39. Re-allocating ONUs to wavelength channels.

can adapt to the actual hot spot distribution. Figure C1.2.40 shows that when the flexible wavelength router is positioned at the second splitting point in the network, the blocking performance is better than the best-case static WDM performance; but more importantly, it is also stable against variations in the hot spot distribution, and thus would allow an operator to guarantee a certain system blocking performance while still optimizing the efficiency of his system's capacity resources. The blocking performance may even be better and stable when positioning the flexible router at the third splitting point; however, this implies that the costs of the router are shared by less ONUs. Locating the router at the second splitting point is a good compromise between adequate improvement of the system blocking performance and system costs per ONU.

C1.2.5.6 Microwave signals over fibre

Fibre-wireless systems may also carry microwave signals directly over fibre. In wireless local area networks (WLANs), the evolution towards larger capacities necessitates higher microwave carrier frequencies. For example, the current IEEE 802.11b WLAN systems transport up to 11 Mbit s^{-1} per carrier in the ISM 2.4 GHz band. The upcoming IEEE 802.11a systems carry up to 54 Mbit s^{-1} per

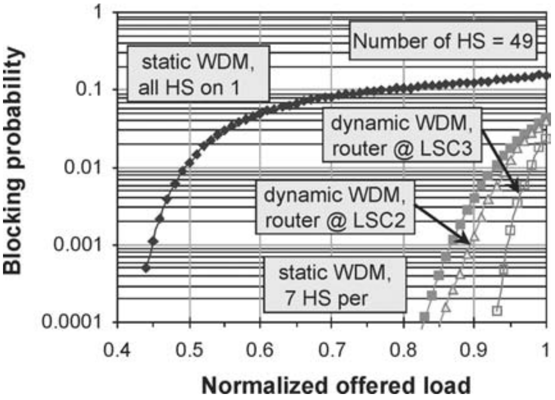


Figure C1.2.40. Improving the system performance by dynamic wavelength allocation.

carrier in the 5.2 GHz band, and 60 GHz systems are under study for providing more than 100 Mbit s^{-1}. With these increasing carrier frequencies, the microwave cells covered by the antenna of a radio access point (RAP) become smaller. Thus more RAPs are needed to serve e.g. all the rooms in an office building, and hence also a more extensive wired network to feed the RAPs. Instead of generating the microwave signals at each RAP individually, feeding the microwave signals from a central headend site to the RAPs enables to simplify the RAPs considerably. The signal processing functions can thus be consolidated at the headend site. Due to its broadband characteristics, optical fibre is an excellent medium to bring the microwave signals to the RAPs.

Carrying multi-gigahertz analogue signals over fibre requires very high frequency optical analog transmitters and receivers, including careful fibre dispersion compensation techniques. An attractive alternative avoiding the transport of multi-gigahertz intensity-modulated signals through the fibre is to apply heterodyning of two optical signals of which the difference in optical frequency (wavelength) corresponds to the microwave frequency. When one of these signals is intensity-modulated with the baseband data to be transported, and the other one is unmodulated, by optical heterodyning at the photodiode in the receiver the electrical microwave difference frequency signal is generated, amplitude-modulated with the data signal. This modulated microwave signal can via a simple amplifier be radiated by an antenna; thus a very simple low-cost radio access point can be realized, while the complicated signal processing is consolidated at the headend station. This approach, however, requires two light sources with narrow spectral linewidth and carefully stabilized difference in optical emission frequency. An alternative approach requiring only a single optical source is shown in figure C1.2.41 [18]. The optical intensity-modulated signal from a laser diode is subsequently intensity-modulated by an external Mach–Zehnder modulator (MZM) which is biased at its inflexion point of the modulation characteristic and driven by a sinusoidal signal at half the microwave frequency. At the MZM's output port a two-tone optical signal emerges, with a tone spacing equal to the microwave frequency. After heterodyning in a photodiode, the desired amplitude-modulated microwave signal is generated. The transmitter may also use multiple laser diodes, and thus a multi-wavelength radio-over-fibre system can be realized with a (tunable) WDM filter to select the desired wavelength radio channel at the antenna site. The system is tolerant to fibre dispersion, and also the laser linewidth is not critical as laser phase noise is largely eliminated in the two-tone detection process.

An alternative approach to generate microwave signals by means of a different kind of remote optical processing, named optical frequency multiplying, is shown in figure C1.2.42 [19]. At the headend station the wavelength λ_0 of a tunable laser diode is swept periodically over a certain range $\Delta\lambda_{sw}$, with a sweep frequency f_{sw}. Alternatively, the wavelength-swept signal can be generated with a continuous-wave operating laser diode followed by an external phase modulator that is driven with

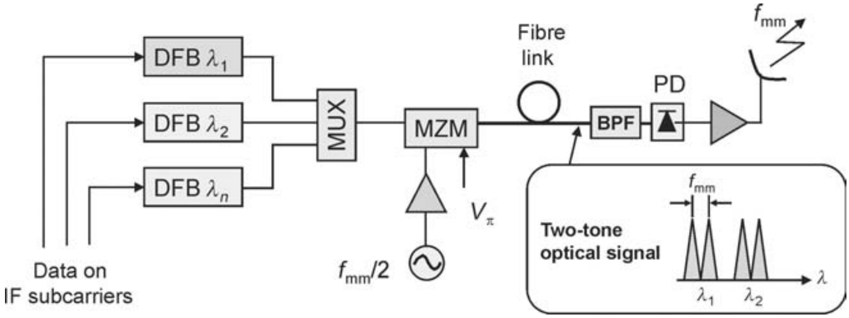


Figure C1.2.41. Generating microwave signals by heterodyning.

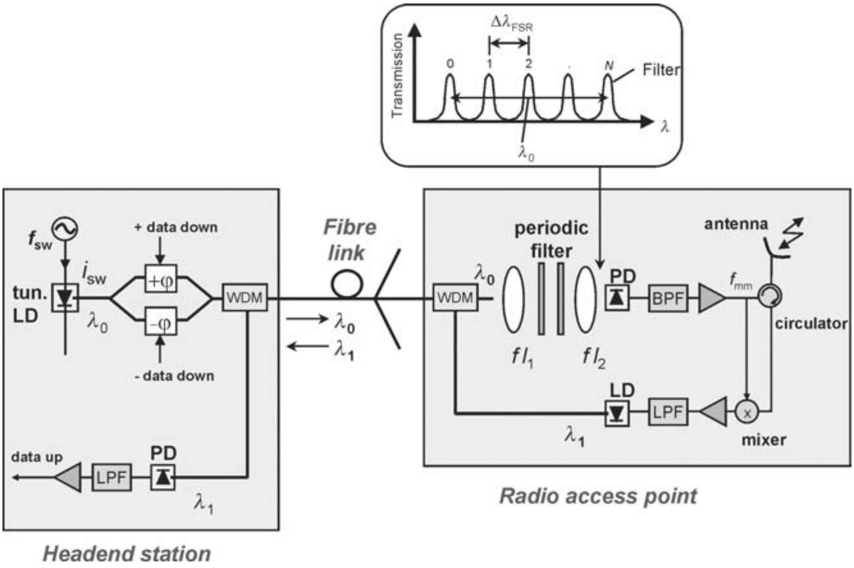

Transmission
$\Delta\lambda_{FSR}$
0
1
2
N
Filter
λ
λ_0
f_{sw}
i_{sw}
tun. LD
λ_0
+ data down
$+\varphi$
$-\varphi$
- data down
WDM
Fibre link
$\rightarrow \lambda_0$
$\leftarrow \lambda_1$
data up
LPF
PD
λ_1
Headend station
periodic filter
WDM
λ_0
fl_1
fl_2
PD
BPF
f_{mm}
antenna
circulator
LD
LPF
mixer
λ_1
Radio access point

Figure C1.2.42. Generating microwave signals by optical frequency multiplying.

the integral of the electrical sweep waveform. The intensity of the wavelength-swept signal is on/off modulated with low frequency chirp by the downstream data in a symmetrically driven MZM. After travelling through the fibre network, the signal transverses at the receiver an optical filter with a periodic bandpass characteristic. When the wavelength of the signal is swept back and forth over N filter transmission peaks, the light intensity impinging on the photodiode fluctuates at a frequency $2Nf_{sw}$. Thus the sweep frequency is multiplied, and a microwave signal with carrier frequency $f_{mm} = 2Nf_{sw}$ plus higher harmonics is obtained. The intensity-modulated data is not affected by this multiplication process, and is maintained as the envelope of the microwave signal. The microwave signal is subsequently transversing an electrical bandpass filter (BPF) to reject the unwanted harmonics. Simulations have shown that the microwave signal is very pure; its linewidth is nearly independent of the linewidth of the tunable laser. The periodic optical bandpass filter can be advantageously implemented by a Fabry–Perot filter with a free spectral range $\Delta\lambda_{FSR}$ which is N times as small as the wavelength sweep range $\Delta\lambda_{sw}$. The microwave signal can also carry more advanced data modulation schemes; e.g. 16-level quadrature amplitude modulated (16-QAM) signals may be modulated on a subcarrier first, and then drive the MZM. The main advantage of this optical frequency multiplying method is that the fibre network is carrying only signals at moderate frequencies (up to the sweep frequency f_{sw}, e.g. up to 1 GHz), while at the antenna site microwave signals with carrier frequencies in the tens of gigahertz region are generated. Thus dispersion problems in the fibre network and other bandwidth limitations in the transmitter are effectively circumvented. The system does not rely on heterodyning, and thus may also operate on multimode fibre networks (such as polymer optical fibre).

The system can also transport upstream data from the antenna station to the headend. When no data are sent downstream, the upstream microwave signal arriving at the antenna can be downconverted with the locally generated unmodulated microwave signal. Thus data can be conveyed bi-directionally in time-division duplex mode.

C1.2.6 Concluding remarks

Optical fibre is now generally recognized to be the most powerful medium for transporting information, due to its very low losses and extremely wide bandwidth. Next to space and time multiplexing, the wavelength dimension offers unprecedented opportunities to extend not only the data traffic transport capacity, but also the traffic routing possibilities in networks.

In core networks, optical amplifiers together with wavelength multiplexing techniques have allowed transoceanic links to be bridged with terabits/second total transport capacity. In terrestrial mesh-shaped core networks, wavelength routing also provides link protection and thus improves network availability. IP packet streams may be very efficiently transported through mesh-shaped networks by wavelength-selective optical bypassing of the electronic processing in the packet routers located in the nodes.

In MANs, wavelength multiplexing techniques enable simultaneous broadband data communication between many nodes. They also improve the network availability by offering alternative routing, which enables fast recovery from link or node failures. Through wavelength add-drop multiplexers, multiple connections can be set up between nodes via a hubbing node, or directly between them using an appropriate wavelength assignment strategy.

In access networks, fibre is penetrating steadily towards the end user. Infrastructure costs are the major nut to crack here. Shared-feeder concepts such as the passive optical network tree-and-branch one greatly reduce the installation costs of the fibre network, and can support various multiple access techniques (a.o. ATM and Ethernet for time multiplexed access). In the past, operators have invested a lot in various last-mile networks (e.g. twisted pair, coaxial cable) to reach their residential customers. In upgrading these networks to higher capacity and larger service variety, fibre can support a wide range of last-mile technologies by hybrid combinations such as fibre-twisted pair, fibre-coax, and fibre-wireless. By means of wavelength multiplexing techniques, the fibre feeder part of such hybrid networks can very flexibly host different operators and service categories. Augmented with wavelength routing, capacity-on-demand can be realized, e.g. for handling hot spots, while respecting quality of service requirements. Carrying microwave radio signals in analog form over the fibre offers clear advantages in the implementation of mobile communication networks; the antenna stations can be considerably simplified and thus reduced in costs, while also the mobility functions can be consolidated in the headend station yielding improved efficiency of the signal processing. By combining radio over fibre techniques with flexible wavelength routing, the mobility offered by wireless communication can be powerfully complemented with the broadband dynamics of optical networks.

C1.2.7 Future prospects

With the ongoing improvements in fibre characteristics and development of novel optical amplifier structures, the wavelength range available for communication will stretch from below 0.8 to beyond 1.6 μm, covering a bandwidth of some 200 THz. Further improvements in signal coding yielding a higher spectral efficiency, in ultra-dense WDM and in ultra-high speed OTDM will enable us to exploit this huge bandwidth, and will push the transport capacity of a single fibre beyond 100 $\mathrm{Tbit\,s^{-1}}$.

Realizing these tremendous link transport capacities over fibre links is of little value if the network nodes cannot keep up with handling the data streams. Present-day electronic routing will be replaced by fast optical processing. All-optical packet routing will provide the ultimate in node throughput, by optically inspecting the header, and by making routing decisions at light speed by means of ultra-fast optical logic. Optical memories will provide the intermediate buffering in the nodes, and electronics may be expelled to the edges of the network, thus providing a fully-optical network path.

The end user is to benefit from all these ultra-broadband communication possibilities. Therefore, the optical fibre will not only reach up to his house, but will penetrate into it as well, benefiting of the low installation cost techniques made possible by deploying e.g. large-core multimode (polymer) optical fibre. It will reach close to his personal area network, which due to the ongoing miniatuarization may consist of a myriad of small wireless power-lean terminals, sensors and actuators. These wireless devices will be incorporated not only in his residential living environment, but also in his clothes, his car, etc. Next to the traditional wired terminals such as the TV, these wireless devices will be connected to the fixed in-home and access network by a myriad of small intelligent antennas. Fibre-radio techniques, augmented with optical routing to accommodate dynamically the hot spots, will provide the best match of the ultimate capacity of fibre with the user freedom of wireless. Even in the wireless domain, optics may penetrate by means of intelligently-steered free-space light beams providing the ultimate in wireless transport capacity.

Which in the far foreseeable future will make the communication world an end-to-end globally transparent one, with nearly unlimited communication capacity for anybody, anytime, anywhere, for any kind of service … the ultimate global crystal ball!

References

[1] Fukuchi K *et al* 2001 10.92 Tb/s (273 × 40 Gb/s) triple-band/ultra-dense WDM optical-repeatered transmission experiment *Proc. of OFC 2001* Post-deadline paper PD24 (Anaheim, CA, Mar. 22, 2001)

[2] Nakazawa M *et al* 2000 1.28 Tbit/s – 70 km OTDM transmission using third- and fourth-order simultaneous dispersion compensation with a phase modulator *Proc ECOC 2000* Post-deadline paper 2.6 (Munich, Sep. 3–7, 2000)

[3] Koonen T, de Waardt H, Jennen J, Verhoosel J, Kant D, de Vos M, van Ardenne A and van Veldhuizen E J 2001 A very high capacity optical fibre network for large-scale antenna constellations: the RETINA project *Proc. NOC 2001* (Ipswich, June 25–29, 2001) pp 165–172

[4] Koonen T, Morthier G, Jennen J, de Waardt H and Demeester P 2001 Optical packet routing in IP-over-WDM networks deploying two-level optical labeling *Proc. ECOC'01* paper Th.L.2.1 (Amsterdam, Sep. 30–Oct. 4, 2001) pp 608–609

[5] Koonen T, Sulur, Tafur Monroy I, Jennen J and de Waardt H 2002 Optical labeling of packets in IP-over-WDM networks *Proc. ECOC'02* (Copenhagen, Sep. 8 – 12, 2002) vol 2, paper 5.5.2

[6] van As H R 2002 Overview of the Evolving Standard IEEE 802.17 Resilient Packet Ring *Proc. NOC 2002* (Darmstadt, June 18–21, 2002) pp 277–284

[7] Koonen A M J and van Veen D T 1997 Estimating error probabilities caused by in-band and out-band crosstalk in multi-wavelength all-optical networks *Proc. IEEE/LEOS Symposium* Chapter Benelux (Eindhoven, Nov. 26, 1997) pp 169–172

[8] Tachikawa Y, Inoue Y, Kawachi M, Takahashi H and Inoue K 1993 Arrayed-waveguide grating add-drop multiplexer with loopback optical paths *Electr. Lett.* **29** 2133–2134

[9] van Veen D T, Krommendijk F N, Sikken B H and Koonen A M J 1999 Impact of OADM architecture design on the performance of ring networks *Proc. ECOC '99* vol II (Nice, Sep. 1999) pp 48–49

[10] Frigo N J, Iannone P P, Reichmann K C, Walker J A, Goossen K W, Arney S C, Murphy E J, Ota Y and Swartz R G 1995 Demonstration of performance-tiered modulators in a WDM PON with a single shared source *Proc. ECOC '95* (Brussels, Sep. 17–21, 1995) pp 441–444

[11] Ueda H, Okada K, Ford B, Mahony G, Hornung S, Faulkner D, Abiven J, Durel S, Ballart R and Erickson J 2001 Deployment status and common technical specifications for a B-PON system *IEEE Commun. Mag.* **39** 134–141

[12] Effenberger F J, Ichibangase H and Yamashita H 2001 Advances in broadband passive optical networking technologies *IEEE Commun. Mag.* **39** 118–124

[13] Schoop R, Fredericx F, Koonen T and Hardalov C 2002 WDM isolation requirements for CATV in BPON *Proc. ECOC 2002* (Copenhagen, Sep. 8–12, 2002)

[14] Vetter P *et al* 2002 Study and demonstration of extensions to the standard FSAN BPON *Proc. ISSLS 2002* (*XIVth International Symposium on Services and Local Access*) (Seoul, Apr. 14–18, 2002) paper 4–2, pp 119–128

[15] van de Voorde I, Martin C, Ringoot E, Slabbinck H, Tassent M, Bouchat C, Goderis D and Vetter P 2000 The super PON demonstrator: a versatile platform to evaluate possible upgrades of the G.983 APON *Proc. ISSLS 2000: XIIIth International Symposium on Services and Local Access* (Stockholm, June 18–23, 2000) paper 14–3, p 10

[16] Koonen T, Muys T, van der Plaats C, Heemstra de Groot S M, Kenter H J H N, Niemegeers I G M M and Slothouber F N C 1997 TOBASCO: An Innovative Approach for Upgrading CATV fibre-coax networks for broadband interactive services *IEEE Commun. Mag.* **35** 76–81

[17] Koonen T, Steenbergen K, Janssen F and Wellen J 2001 Flexible reconfigurable fiber-wireless network using wavelength routing techniques: the ACTS project AC349 PRISMA *Photon. Network Commun.* **3** 297–306

[18] Griffin R A, Lane P M and O'Reilly J J 1999 Radio-over-fibre distribution using an optical millimeter-wave/DWDM overlay *Proc. OFC '99* (San Diego, Feb. 22–25, 1999) paper WD6

[19] Koonen T, Ng'oma A, Smulders P, van den Boom H, Tafur Monroy I, van Bennekom P, Khoe G-D 2002 In-house networks using polymer optical fibre for broadband wireless applications *Proc. ISSLS 2002: XIVth International Symposium on Services and Local Access* (Seoul, Apr. 14–18, 2002) paper 9–3, pp 285–294

Koonen T, Ng'oma A, Smulders P, van den Boom H, Tafur Monroy I and Khoe G-D 2003 In-house networks using multimode polymer optical fiber for broadband wireless services *Photon-Network Commun.* **5** 177–187

[20] Ryf R, 2002 Optical MEMS for optical networking *Proc. ECOC'02* (Copenhagen, Sep. 8–12, 2002) vol 1, Tutorial 2

[21] Ryf R *et al* 2002 Multi-service optical node based on low-loss MEMS optical crossconnect switch *Proc. OFC 2002* (Anaheim, CA, Mar. 17–22, 2002) paper ThE3, pp 410–411

C1.3
Optical switching and multiplexed architectures

Dominique Chiaroni

C1.3.1 Introduction

This chapter gives a positioning of optical switching technologies in the next generation of systems and networks. After giving two introduction scenarios, one for the metro part and one for the backbone part, this chapter addresses feasibility issues. For the metro part, three introduction scenarios are presented. The first one exploits the well known circuit switching techniques that can be introduced rapidly on the market. The second one proposes a packet switching technique to have a better bandwidth exploitation. Finally, with the emergence of new optical functions/devices, the third scenario describes how it is possible to propose a full flexible packet ring network that is really competitive with respect to other electronic alternatives. For the backbone part, the first introduction will be probably in the core of large routers, competing with current smart routers (router + cross-connects). The second introduction is for a new network concept, being disruptive with what exists but pushing towards a transparent compliant with a multiservice environment and fully evolutive in capacity. Finally, the last scenario describes how it is possible to go into an all-optical approach through the description of key optical functions required to make this concept realistic at a lower cost.

With the massive penetration of the Internet protocol in the network, the telecommunication domain has turned a new corner. The broadband access to this new technology, opening the way to many residential applications, creates a revolution for the next generation of switching systems. The first revolution is the traffic volume. Personal computers becoming more and more powerful are generating a traffic through files that could not be envisaged even two years before. The second revolution is probably the traffic profile evolution, moving from constant bit rate to variable bit rate, always driven by personal computer capabilities (video applications, HDTV, net-shopping, net-courses, games, etc).

Optical technologies could appear in the next four years as an important technology to grow the capacity of systems while preserving the simplicity, the reliability and the performance of the systems. But more importantly, optical packet switching technologies could become efficient techniques to really fit with the statistical behaviour of the traffic profile to preserve bandwidth utilization as much as possible. One of the key issues in such packet-switched networks is the identification of the best packet format (variable packets or fixed packets). Several European projects have concentrated their efforts on this important topic like the RACE 2039 ATMOS project, the ACTS 043 KEOPS project and more recently the IST DAVID project.

Thus in this chapter, after a positioning of optical switching in the next generation of systems and networks, the benefits of multiplexed architectures will be presented. Solutions for a progressive introduction of this technology in the metro are described, highlighting the required technology and addressing physical feasibility as well as performance issues. Opportunities for the backbone are also presented with the objective of highlighting the most promising approaches. For a pragmatic approach, criteria introducing this technology on the market are listed, but more importantly, a basic cost approach leading to the winning solution is mentioned. Finally, a conclusion is drawn.

C1.3.2 Positioning optical switching techniques in the next generation of systems and networks

In this section, we will present the advantage of optics with respect to electronics, but more importantly, how optics can be exploited to complement electronic technology to really make the most of both technologies.

C1.3.2.1 Why optical switching?

To give some arguments we must list the advantages and the drawbacks of optics.

Main advantages of optics:

- Low power consumption: as an example a laser exploited in direct modulation. A laser operating at 622 Mbit s^{-1} or at 2.5 Gbit s^{-1} will be electrically modulated with the same electrical modulation amplitude.
- High reliability: for passive devices it is evident, for active devices (lasers, EDFAs, semiconductor optical amplifiers (SOAs), etc) the reliability is quite high since we exploit a carrier density dynamic, characteristic of the material used.
- Good mode adaptation: The default on the coupling simply introduces losses and the reflection can be easily managed by exploiting tilts.
- Low power dissipation: the interaction photon–photon dissipates less energy than electron–electron, mainly because of the mass.
- High bit rate compliant: a passive guide is *a priori* a high bandwidth medium capable of supporting several terabits of capacity.
- Management of large granularity: The switching can be done at the wavelength level but also at the waveband (group of wavelengths) level.

Main drawback of optics:

- Slow progress in integration: Still a debate between monolithic and hybrid integration.
- Slow progress in low cost packaging: the coupling between passive and active guides still remains a costly technique.
- Polarization sensitive: The characteristic of some materials often depends on the polarization state of the light.

In summary, optics is very interesting when the switching granularity is high, exploiting the wavelength division multiplexing (WDM) dimension to make simple structures. In electronics we need to demultiplex at the wavelength level and then at the bit rate level, in optics we simply need one device.

To switch at the WDM granularity, we have commercially available devices like opto-mechanical switches, thermo-optical switches, electro-optical switches, MEMs, etc for slow switching applications and for fast switching SOAs, digital optical switches, etc.

C1.3.2.2 Granularities of switching

In optics, we can switch wavebands (group of wavelengths), wavelengths or optical packets.

Switching of wavebands

The switching of wavebands is particularly interesting in the following cases:

- Inside systems to reduce the number of switching elements. This is the case for a large number of optical cross-connects or switches.
- At the network level when the traffic is aggregated enough to tolerate a waveband switching with a good bandwidth utilization. This is the case for pipes bridging networks and where the traffic matrix is quite stable.

The waveband switching really exploits the potential of optics, since it reduces the complexity of the switching process with respect to electronic techniques. This technique has to be exploited as much as possible to make a system or a network concept really competitive to electronic techniques but only when the traffic matrix is stable enough not to penalize the average load of the waveband.

Switching of wavelengths

The switching of wavelengths is particularly interesting as follows:

- Inside optical systems, to switch at the line bit rate without time demultiplexing, which is an advantage with respect to electronic techniques, especially when the bit rate is high ($\geq 10\,\mathrm{Gbit\,s^{-1}}$). Generally, the wavelength switching is used to relax the power budget to offer higher switching capacity. In fact, with one important parameter being the optical signal to noise ratio, it is fundamental to preserve it to be able to address the higher throughputs. This is generally imposed when input powers launched into optical amplifiers cannot exceed a certain value. By this way, as the optical signal to noise ratio is always a function of the channel input power, if the total input power is the channel power, then we reach the maximum optical signal to noise ratio value in the architecture. This is the case for a major part of the optical cross-connects or switches.
- At the network level when the traffic is aggregated enough and stable enough not to cope with traffic transient effects. This technique is still efficient when it is possible. The switching is done at the line bit rate. It is more efficient for backbones than for metro networks, for example, simply because of the traffic matrix characteristic.

Switching of optical packets

The switching of optical packets can be processed at the wavelength level or at the waveband level. The only difference with respect to wavelength or waveband switching is that the ON state is relatively short, in the order of a few 100 ns or μs. The switching of optical packets is particularly interesting as follows:

- Inside optical switching systems to create a datagram connection or a virtual circuit connection. The switching can be done at the wavelength level (classical optical packet switching) or at the waveband level (we define then a WDM packet) for a short time. Like for the previous cases, the WDM dimension is preferred where possible to reduce the number of active components in the architecture considered.

- At the network level, when the traffic profile is sporadic, we have to cope with time constants that cannot fit with seconds or minutes, and where the packet connection is the only one realistic to exploit efficiently the available bandwidth. This is the case for metro networks in particular, but also for backbones if the application bit rate is growing at bit rate.

C1.3.2.3 Optical packet switching: an interesting approach

Context

Among the different switching techniques, the optical packet switching technique is probably the most promising technique for the next generation of networks. The main indicator is the natural evolution of the traffic profile versus packet techniques. Driven mainly by the Internet protocol, we need to cope more with a traffic profile than a traffic matrix as could have been the case in the past. The main reason is the drastic change of telecommunication applications moving from telephony to data. In addition, the rapid introduction of PCs at home as multimedia machines, pushes telecom companies to find solutions to offer a higher quality of service at a lower cost. This new form of traffic imposes new infrastructures capable of handling the required capacity and to provide at the same time the required flexibility to offer low cost connections.

Why optical packet switching?

When analysing the traffic profile at the output of a LAN, the sporadic behaviour of the traffic, often modelled with self-similar functions, clearly points to the problem. We then need to adapt the network concepts to the traffic nature coming from the access. And how do we adapt such a variability of the traffic profile with circuit connections while having a good efficiency? The answer lies in the high aggregation level which requires a grouping of different LANs, and this is not always possible. Another solution is to cut the circuit into pieces (packets) trying to follow the traffic evolution at a scale comparable to the scale of the incoming traffic. Even if the technique is more complex to manage, it is undoubtedly the most efficient way to optimize the bandwidth utilization, more in the time domain today, than in the volume domain in the recent past.

What kind of packets: fixed packet or variable packets?

There is still a debate on the choice of the packet size, and for main arguments in favour of the variable packet: the optical packet size always follows the incoming packet size, and for the fixed packet: the optical packet format contributes to a better management of the performance. In both cases, arguments are acceptable but the reality is more complex.

It is evident to say that where the contention can be managed easily (a case of small or simple topologies), the variable packet has to be envisaged. But in the case of a large topology (meshed or other), the resolution of the contention is then local in each node and the control of the traffic profile inside the network becomes fundamental. In that case the fixed packet format is the only reasonable format. Another alternative, probably the best, is the adoption of a concatenated packet. For best effort, the concatenation is created to really follow the incoming packet profile. For high priority traffic where the delay is fundamental, small packets will always experience the smallest delay in the network. In this way the technique can be adapted to a multiservice environment (figure C1.3.1).

Figure C1.3.2 gives a comparison between fixed, variable and concatenated packets.

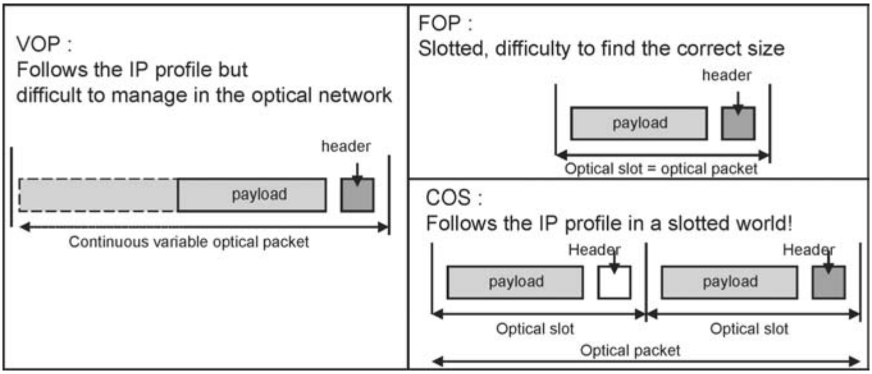

VOP :
Follows the IP profile but
difficult to manage in the optical network
header
payload
Continuous variable optical packet
FOP :
Slotted, difficulty to find the correct size
header
payload
Optical slot = optical packet
COS :
Follows the IP profile in a slotted world!
Header
payload
Header
payload
Optical slot
Optical slot
Optical packet

Figure C1.3.1. Different optical packet types that can be adopted.

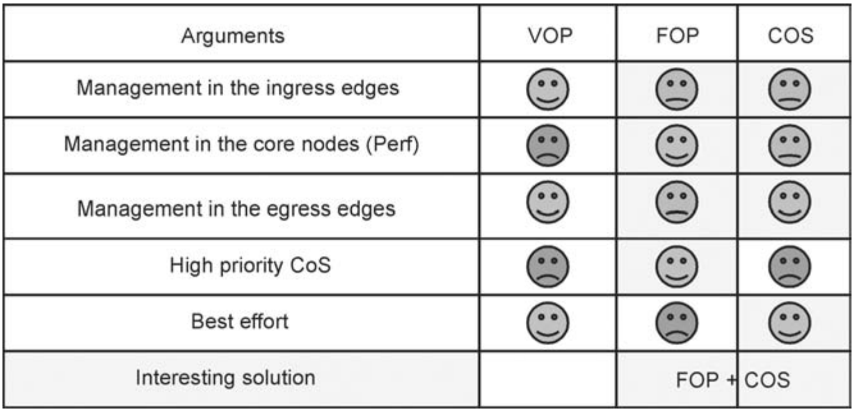
C1.3.3 Benefits of multiplexed architectures

C1.3.3.1 The WDM dimension, a dimension not only for transmission systems

The WDM technique is often assimilated to the transmission domain. But, in fact, this technique is very useful in optics for many purposes.

The WDM technique can be exploited for the following objectives:

- To increase the capacity of a link without increasing the TDM bit rate at values difficult to manage. This first application was found rapidly. But there was also the requirement to reduce the number of active components. As a good example, the EDFA amplifying a group of wavelengths has rapidly replaced the classical single-channel repeater.
- To facilitate the optical processing to avoid the interferometric noise when interleaving several pulses at high bit rate. The technique is used in OTDM, in coding format techniques, in multiplexing techniques (bit level or packet level), etc.



Arguments	VOP	FOP	COS
Management in the ingress edges	🙂	🙁	🙁
Management in the core nodes (Perf)	☹	🙂	🙁
Management in the egress edges	🙂	🙁	🙂
High priority CoS	☹	🙂	☹
Best effort	🙂	☹	🙂
Interesting solution		FOP + COS	

Figure C1.3.2. Comparison: variable optical packets *versus* fixed optical packets *versus* concatenated optical slots.

- To ease the contention resolution. In fact the wavelength becomes a new support, with different colours, and the colour is selected in real time to avoid collisions. This technique is particularly used in optical packet switching network concepts.
- In switching to switch high granularities. It can be introduced to reduce the number of components in an optical architecture, or in cases of low connectivity while addressing high throughputs.

C1.3.3.2 Benefits of the WDM dimension in multiplexed architectures

In multiplexed architectures, the WDM dimension is exploited for different purposes.

In the following, we will describe where the WDM dimension can be exploited efficiently, and what kind of benefit we can expect.

In metropolitan networks

In the case of optical rings, the WDM dimension can be exploited first to provide an upgradability of the network in terms of allocated resources, simply by attributing progressively bands of wavelengths. The advantage of the band is mainly to relax the filtering constraints during the cascade of several nodes. This advantage was raised many times in circuit switching platforms. If we want to exploit the same WDM infrastructure while introducing packet techniques, the notion of band can then be advantageously exploited to reduce the latency in the transmitting parts. In fact we can exploit the statistical time multiplexing of packets over a group of wavelengths to increase the chance to insert a packet in the line when we can have access to a group of wavelengths instead of one.

In summary, we can see that for optical rings, the WDM dimension can relax physical constraints and, in addition, improve the performance in packet rings in terms of latency.

To illustrate the benefit of the WDM dimension, the IST DAVID project is proposing a multi-ring optical packet MAN exploiting the WDM dimension for these two main aspects.

In backbone networks

In backbone networks where the topology is generally meshed, the WDM dimension is exploited for four main reasons:

- Cost and compactness reasons. In optical cross-connect/packet switching architectures, the WDM dimension can be exploited to reduce the number of switching components to make the architecture compact and low cost. The objective is to switch in the WDM dimension as much as possible in the limit of the required power budget. In cross-connect architectures, the dynamic management of the wavelength dimension can, in addition, contribute to a drastic reduction of the number of interconnected fibres.
- Power consumption reasons. By exploiting the WDM dimension, the processing is done at the WDM dimension. In electronics, a double demultiplexing is required: at the wavelength level, and at the bit rate level. This is, for example, the case for coarse synchronization stages where the WDM dimension can be exploited efficiently at the WDM level to reduce the complexity of many electronic structures.
- Physical reasons. The number of channels switched can also vary inside the architecture to preserve the optical signal-to-noise ratio at the output. For example, in the first stage of an optical architecture,

the WDM switching can be done on a large number of channels (8 or 16), and then progressively reduced to a lower number, converging then to one channel switched.

- Performance reasons. In packet architectures, the WDM dimension can be exploited, to reduce the packet loss rate and the latency at the same time. When the packets can be addressed onto different wavelengths, the statistical multiplexing on these wavelengths let a freedom degree for the choice of the wavelength at the output of an optical packet switch. By this way, the contention can be avoided simply by re-affecting a new wavelength to a packet instead of putting it in a queue. The benefit is double since the packet loss rate can be improved, with the same amount of digital memory and the latency is preserved, since the packet will not experience a queue.

C1.3.4 Specificity of the metro and proposed solutions

C1.3.4.1 Introduction scenario

To take a pragmatic approach, it is fundamental to identify what could be the introduction scenario of a technology at the time. These paragraphs intend to position the optical switching technology for the metro in a time scale.

Before developing these scenarios, it is also important to list the specificity of the metro part.

In the metro, it is clear that the cost is probably the most important parameter together with the performance. The cost addresses the hardware part but also the means adopted to exploit the bandwidth in the best way. Due to the traffic profile coming from the access, the important point is really to preserve the bandwidth. Thus packet techniques will be preferred in this part of the network.

What about the WDM dimension?

The WDM dimension is expensive in the metro but there are also some arguments in favour of the introduction of the WDM in this part of the network.

Due to the current traffic profile, not constraining the bandwidth utilization too much (because the native bit rate is still quite low), a circuit platform is interesting. To reduce the cost of the WDM dimension while having enough resources to cope with the traffic volume and profile, probably the waveband approach is the best one. It guarantees upgradability (sub-band per sub-band), physical performance (relaxing filtering constraints), and simplicity (circuit switching) with an existing infrastructure (fibres already installed). But if the native bit rate increases together with the variance, there will be a need to go into a lower granularity to have a better utilization of the optical bandwidth. Then the optical packet technique could be easily introduced making use of an existing infrastructure. The gap becomes natural.

Thus in the following presentation, we will introduce the circuit switching technique as a first step with a progressive migration towards optical packet switching techniques, which lead to a really efficient platform.

Short term introduction: optical circuit switching technology

Due to the current traffic profile, optical circuit switching could be rapidly introduced in the market for many reasons:

- The optical technology is mature enough and commercially available to envisage its utilization in optical platforms. It requires active devices like ILMs, receivers and basic passive devices (demultiplexers, couplers, filters, etc).

- It is a very simple technique that can be implemented rapidly.
- The optical technology can exploit advantageously an existing WDM infrastructure. Currently, all the fibres installed are not exploited.
- The management of such a network is mature enough to propose products. Management studies have been carried out leading to clear information models, protection scenario and well-defined monitoring techniques.
- The traffic profile is not yet so critical to envisage using such a technique in the metro. The bit rate at the output of PC is sufficiently low to have enough aggregation level at the output of the LANs.
- This kind of platform can be easily upgraded with optical packet techniques to be fully compatible with the future traffic profile driven by a powerful PC and FFTH. This is also a strong argument for operators who want to invest with the possibility to upgrade their platform according to traffic constraints.

Examples of implementation are ring topologies, star topologies, and mesh topologies.

In ring topologies, we can find the single ring or the multi-ring approach. Figures C1.3.3(*a*) and C1.3.3(*b*) illustrate three interesting topologies:

The particularity of this approach is in the optical add/drop multiplexing structure.

We can list mainly:

- passive optical add/drop multiplexers (OADM) and
- dynamic optical add/drop multiplexers.

The passive structure can be used when the traffic matrix is very stable whereas the dynamic structure can allow some adaptation of the allocated resources to follow at least the envelope of the traffic matrix. This last type of active structure is particularly interesting when the time constants are in the range of a few hours.

Figures C1.3.4(*a*) and C1.3.4(*b*) describe the two main structures of OADMs.

In the metro part, the WDM granularity for the upgradation of the network capacity will depend on the cost of the intervention. And in some cases the upgradation of one wavelength is not the most cost saving solution. For that purpose, it is important to identify the minimum WDM granularity for the upgradation. This minimum granularity, which can be in the order of few wavelengths (2 or 4), can impact dramatically the network infrastructure. If we take into account the physical limitations in the cascade of filters, it is clear that the sub-band approach is an interesting approach. Figure C1.3.5 gives an overview of a ring MAN adopting a sub-band strategy.

Medium term introduction: optical packet switching technology

The optical packet technology could be introduced as a required second step simply to face the traffic profile evolution. In that case, a packet technique will be required on the basis of the infrastructure already installed. The upgradation is made simply by changing the optical ring access node. Two new sub-blocks are then mandatory: the opto-electronic part and the electronic interface compliant with different classes of services.

The more pragmatic approach is the adoption of a commercially available technology. Several concepts are proposed, all based on the adoption of ILMs and SONET-like receivers. The RPR concept is probably the most representative one.

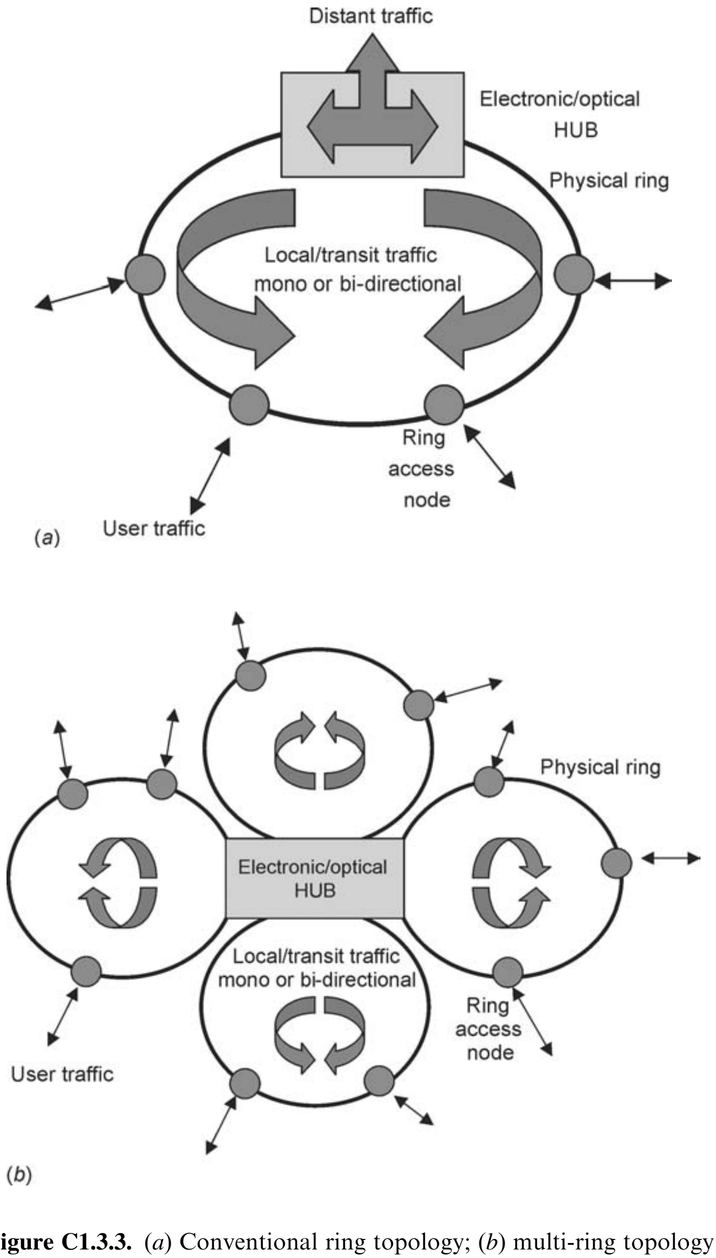

Distant traffic
Electronic/optical HUB
Physical ring
Local/transit traffic mono or bi-directional
Ring access node
User traffic
(a)
Physical ring
Electronic/optical HUB
Local/transit traffic mono or bi-directional
Ring access node
User traffic
(b)

Figure C1.3.3. (*a*) Conventional ring topology; (*b*) multi-ring topology.

Medium term introduction: full flexible optical packet switching technology

In the case of long-term approaches, the adoption of an advanced technology is then mandatory.

In particular, two important components are: a fast tuneable source and a fast wavelength selector. These two components have already been demonstrated feasible (e.g. Agility for the tuneable sources, NTT for the wavelength selectors).

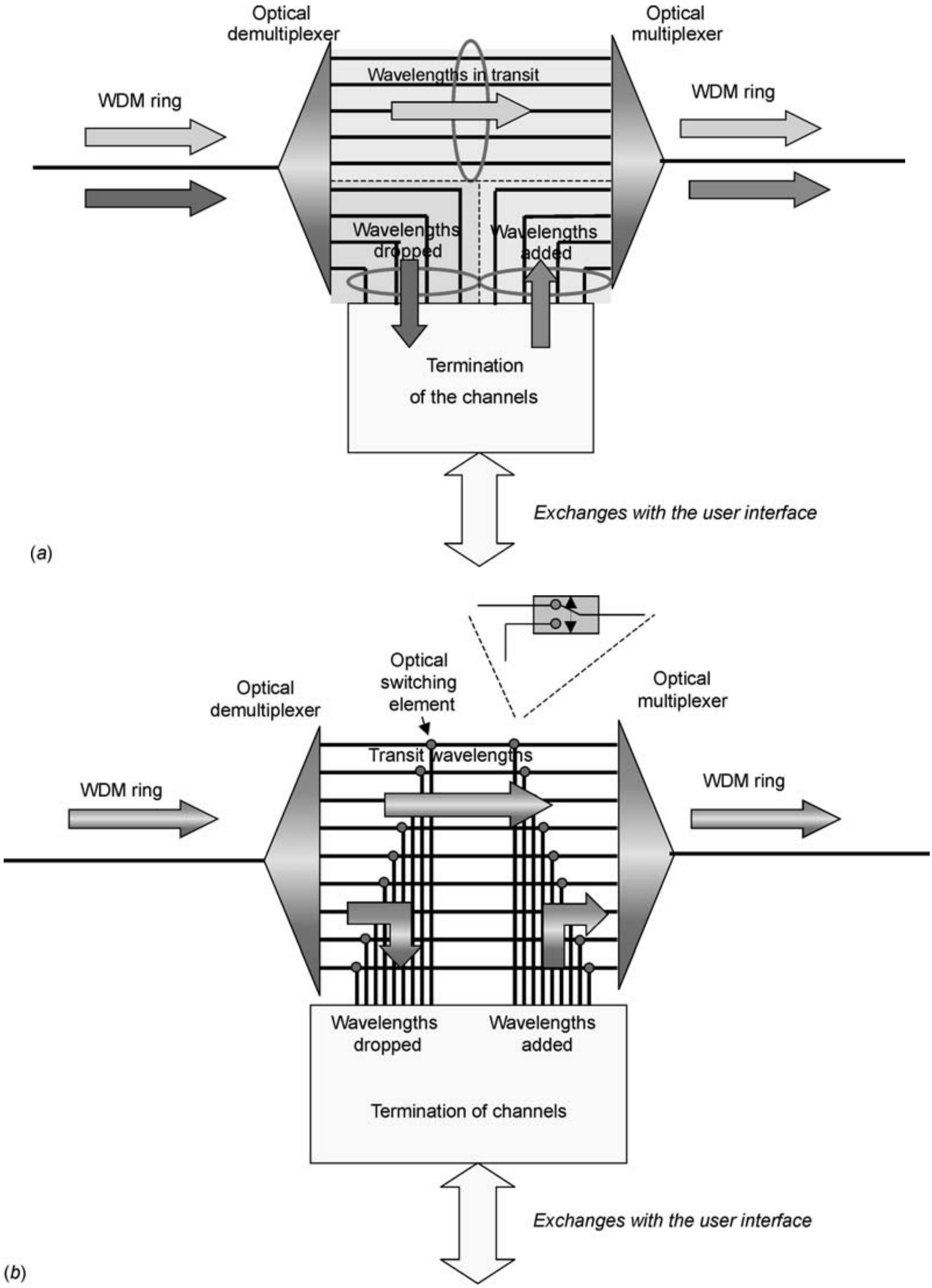

Optical demultiplexer
Optical multiplexer
Wavelengths in transit
WDM ring
WDM ring
Wavelengths dropped
Wavelengths added
Termination of the channels
Exchanges with the user interface
(a)
Optical switching element
Optical demultiplexer
Optical multiplexer
Transit wavelengths
WDM ring
WDM ring
Wavelengths dropped
Wavelengths added
Termination of channels
Exchanges with the user interface
(b)

Figure C1.3.4. (*a*) Fixed optical add/drop multiplexer. (*b*) Dynamic optical add/drop multiplexer: an optical switching element guarantees the dynamic behaviour.

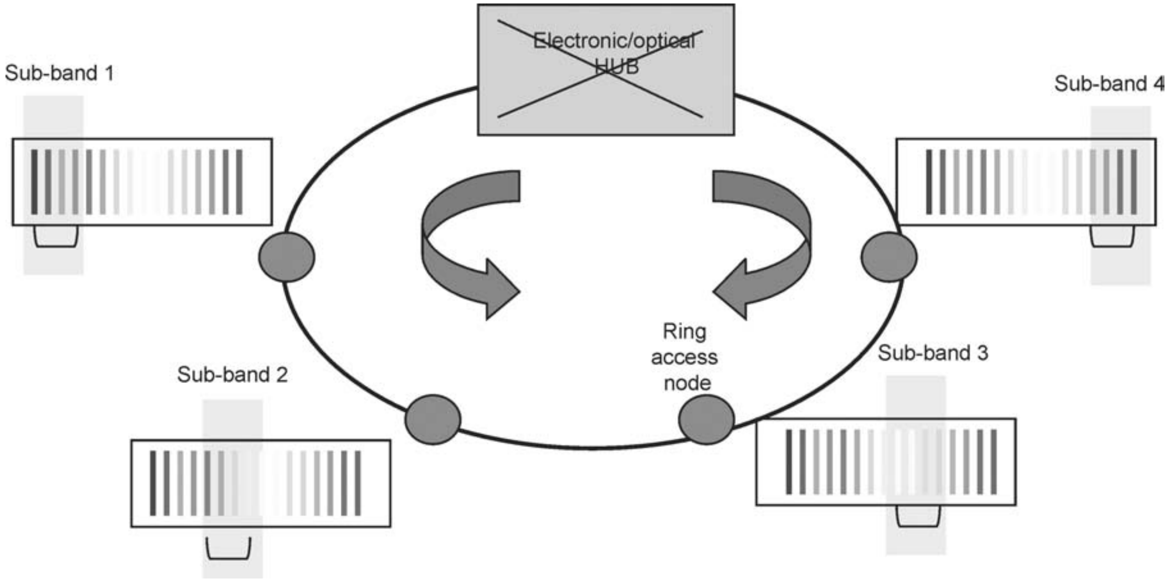


Figure C1.3.5. Sub-band introduction in an optical ring.

These components will be responsible for:

- at the transmission side (tuneable laser): limitation of the latency in the output queue by exploiting the WDM dimension;
- at the receiving side (wavelength selector): the guarantee of the wavelength transparency to receive any packet from the optical bandwidth;
- at the transit side (wavelength selector): the guarantee to introduce some fairness mechanisms when a high priority traffic with a minimum of latency is required.

The structure of the optical add/drop multiplexer is depicted in figures C1.3.6(*a*) and C1.3.6(*b*).

C1.3.4.2 Technology identification and feasibility

The technology required to introduce these concepts can be split into two categories:

- A first category based on commercial devices.
- A second category based on an advanced technology.

In the following, we will list the required commercial devices, and we will describe in more detail potential new advanced technologies that open the way to really attractive system functionalities.

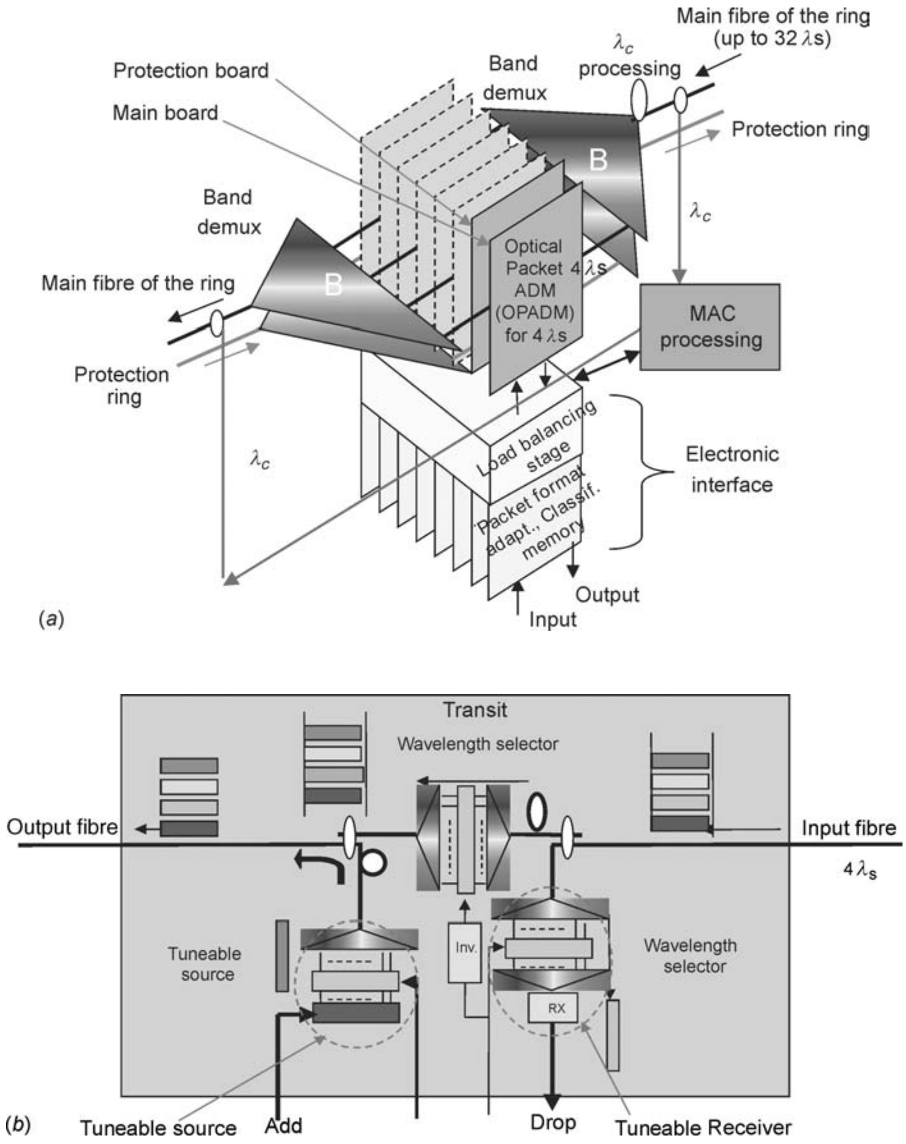

Main fibre of the ring (up to 32 λs)
λc processing
Band demux
Protection board
Main board
Protection ring
B
λc
Band demux
Optical Packet 4 λs ADM (OPADM) for 4 λs
MAC processing
Main fibre of the ring
B
Protection ring
Load balancing stage
Packet format adapt., Classif. memory
Electronic interface
λc
Output
Input
(a)
Transit
Wavelength selector
Output fibre
Input fibre
4 λs
Tuneable source
Inv.
Wavelength selector
RX
(b)
Tuneable source
Add
Drop
Tuneable Receiver

Figure C1.3.6. (*a*) Structure of the ring access node for an upgrade of the circuit switching platform into a packet switching platform. (*b*) Structure of the optical add/drop multiplexer including the required functionality to fully manage the transit traffic.

Commercially available technology

We can classify two types of devices: passive devices and active devices.

- For passive devices we will need couplers, attenuators, isolators, fixed filters, circulators, interconnection fibres, connectors, (de)multiplexers, etc. For filters and (de)multiplexers we can have different specifications if we are addressing channel filtering or band of channels filtering.

In all the cases, these components are products and today fit all the required specifications for system applications.

- For active devices, we will need lasers, ILMs, slow tuneable lasers, optical amplifiers, photodiodes, etc. Particular interest must be devoted to optical amplifiers since we can distinguish two types of optical amplifiers: fibre based amplifiers and semiconductor amplifiers. For fibre based amplifiers, largely introduced in point-to-point transmission systems, the main preoccupation is to find the best way to make these devices very cost effective. For semiconductor optical amplifiers, even if they are on the market today, they currently suffer from a small market. However, the generic potential of such a component for different functions, like optical gating or optical conversion, makes this component a promising one for future applications.

Advanced technology and feasibility issues

This is the most promising technology to really propose something new and disruptive with respect to what exists today.

In the following, we will illustrate the four main components:

- the SOA for gating functions;
- the tuneable source;
- the wavelength selector, a strategic fast tuneable filter; and
- the packet mode receiver.

SOA for gating

A SOA is basically a laser where the facets have been treated in order to eliminate the resonant cavity. Only the amplification medium is exploited. To be used as an optical gate, the SOA requires a high frequency driver interconnected to the SOA. The driver sends a control signal which can be forced at the ON or OFF state. Because of the short carrier lifetime, a SOA can be switched with response times in the order of few tens of nanoseconds. At present, this component is exploited by many laboratories and has been demonstrated to be feasible for many applications.

There are roughly three types of SOA: SOA with high confinement factor or long active section to achieve wavelength conversion at high bit rate, SOA with low confinement factor or short active section to exploit more the linear characteristic of the gain, and SOA with an internal clamping to have a strictly linear response not to create distortion on the signal crossing the component.

As for fibre amplifiers, there are three classes of SOAs: pre-amplifiers, in line amplifiers and boosters.

In the metro area, and according to the two network models described previously, three structures of SOAs are particularly interesting:

SOA gate at the output of an ILM

An SOA interconnected at the output of a modulated source is the basic schematic we can envisage.

The advantage of this solution is that it is commercially available today. The main drawback is that there is a need to have control of the cross-gain modulation. One interesting solution is the use of a clamped-gain SOA to avoid any cross-gain modulation. The SOA used only in its linear characteristic will provide enough gain to guarantee a sufficiently high ON/OFF ratio with no degradation of the pulse shape. This solution is currently studied in different laboratories to analyse network concepts based on a packet transmission.

Tuneable source using a hybrid integration of a SOA gate array

The SOA gate array is located in front of a laser array. At the output a multiplexer and an integrated modulator are interconnected. The SOAs see only continuous waves and can be switched, thus selecting the wavelength that must be transmitted. The SOAs are in a stable regime, since the input power is a constant. So they do not experience any cross gain modulation as could be the case in the previous use. The preservation of the signal quality (no degradation of the extinction ratio, and no distortion of the bits) makes this SOA array an important device for the building of hybrid tuneable sources.

Figure C1.3.7(*a*) illustrates the structure of a tuneable source based on a gate array, and figure C1.3.7(*b*) shows an integrated 4 gate-array (OPTO + realization).

Tuneable switching sources based on a sampled grating—distributed Bragg reflector structure

Another solution to build a fast tuneable source is to use a SG-DBR laser while integrating a SOA section and an electro-absorption section. The following schematic illustrates the structure of the source. As for the hybrid tuneable source, the SOA sees only a continuous wave which again prevents cross gain modulation. The structure is currently studied by Agility (figure C1.3.8).

Wavelength selector

The wavelength selector is probably one of the other key devices, since it can be comparable to a fast tuneable filter. The principle of operation is very simple. A first demultiplexer demultiplexes the wavelengths, then each wavelength is selected or not, depending on the orders coming from the control part, and finally an output multiplexer regroups the wavelengths selected and contributes to reject the wideband amplified spontaneous emission coming from the SOAs. In principle, only one wavelength is

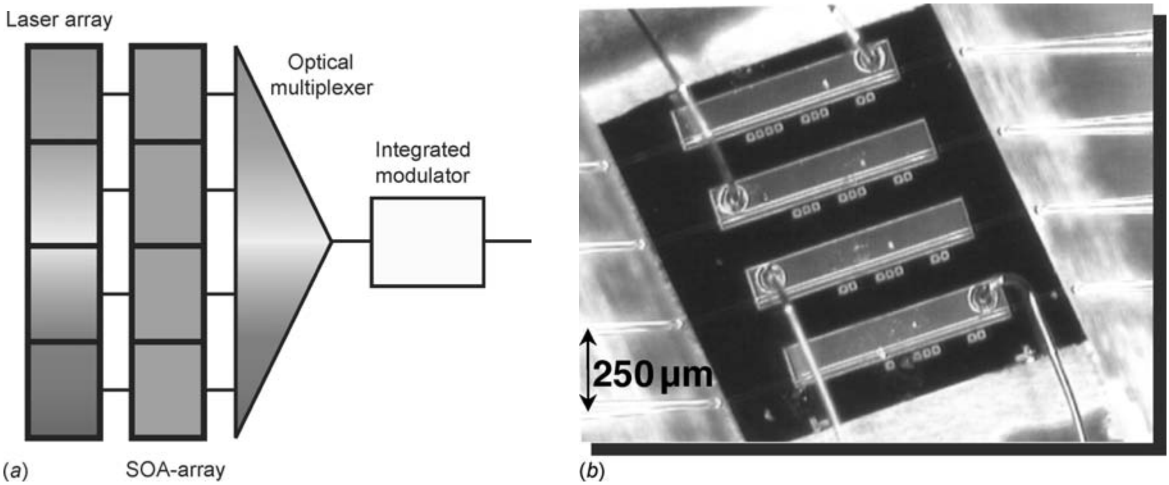


Figure C1.3.7. (*a*) Structure of a hybrid tuneable source using a SOA gate array. (*b*) Photo of a 4 gate-array, key building block of the hybrid tuneable source.

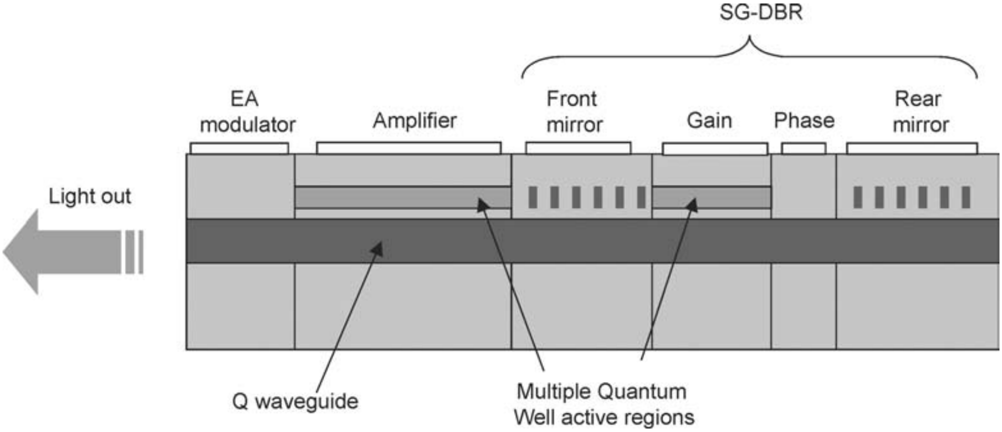

SG-DBR
EA modulator
Amplifier
Front mirror
Gain
Phase
Rear mirror
Light out
Q waveguide
Multiple Quantum Well active regions

Figure C1.3.8. Schematic of an integrated tuneable source integrating a SOA section for amplification of the signal or optical gating.

selected among a group in a normal scenario. But in the case of the network scenario 3 depicted above, the number of output wavelengths can vary from 1 to *N* (*N* being the total number of wavelength at the input of the device).

Figure C1.3.9(*a*) shows the basic structure of a wavelength selector.

Packet mode receiver

In the case of scenario 3, another important technology is required, not at the optical level but more at the electronic level: the packet mode receiver.

The packet mode receiver has to experience a continuous or noncontinuous packet stream exhibiting different packet phases (when aligned to a common reference clock) and suffering from a packet power dispersion. Such kind of receivers are currently studied in different laboratories. We can cite NTT, NEC, Lucent, Alcatel, etc.

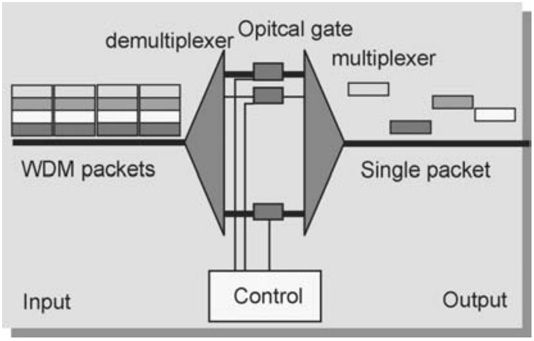

demultiplexer
Opitcal gate
multiplexer
WDM packets
Single packet
Input
Control
Output

Figure C1.3.9. Structure of a wavelength selector.

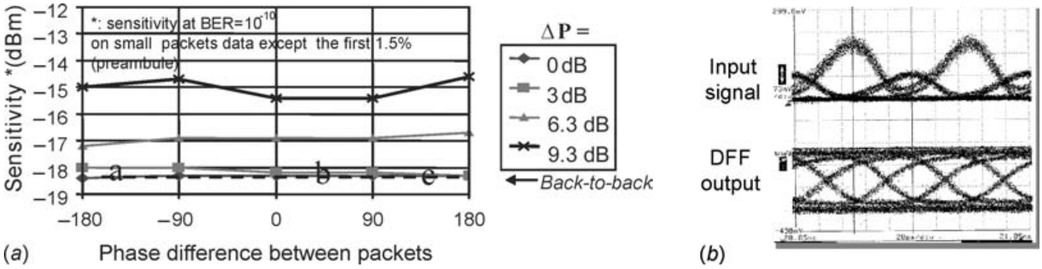


Figure C1.3.10. (*a*) Characteristics of a packet mode receiver. Large power dynamic ranges and fully transparent to the packet phase fluctuation. (*b*) Eye diagram recorded: before and after the 10 Gbit s^{-1} packet mode receiver. The phase is preserved and the amplitude completely equalized.

Currently, this technology has been demonstrated to be feasible for use in different coding formats: Manchester but also RZ or NRZ.

Figures C1.3.10(*a*) and (*b*) show the performance obtained with a packet mode receiver operating at 10 Gbit s^{-1}.

C1.3.4.3 Performance expected for the packet approach

The performance will depend on the packet format adopted.

In the case of variable packets, the contention resolution becomes a bi-dimensional problem: to be able to insert a packet in the ring we have to check if we have a void and if the void is large enough to insert the packet.

In the case of fixed packets, we need to also experience two dimensions, but in that case they are not co-joined. The first problem to solve is the packet filling ratio, imposing some time out to have a good efficiency, whereas the second problem is in the add part: check if there is a free slot.

So the problem is not the same for both cases, and the impact on the performance is not the same.

In the case of variable packets, the main problem to solve will be the management of the voids in the ring so that the last ring access node in cascade is not blocked.

In the case of fixed packets, the main problem to solve is the choice of the good packet size to be compatible with the incoming traffic profile.

So concerning the performance, the variable packet format exhibits a poorer performance compared to the fixed packet. This is the reason why the concatenated approach is probably the optimum solution, providing performance and reliability (packet rhythm present in the ring to ease monitoring aspects).

C1.3.5 Opportunities for the backbone

C1.3.5.1 Introduction scenario

In this part, the objective is to draw a progressive introduction of optical switching techniques for the backbone. For the short term we will propose what we call a high throughput router, combination of a router and a cross-connect, and what we call a high throughput router. Both approaches are of prime importance because they correspond to two reality cases, and two classes of products required to cover different network specificity. As a medium term approach, we will present a multiservice O/E/O network concept based on new features. Finally, we will describe how an all-optical packet switching network could become a reality in the longer term because of attractive features.

Short term introduction: smart router (router + cross-connect) versus multi-terabit class routers/switches

Where smart routers and where high throughput routers?

Smart routers are required for the dorsal network and where the aggregation is forced by the poor connectivity of the nodes and by the huge amount of traffic that must be transported. This type of product is particularly interesting when the traffic matrix is stable enough to make semi-permanent connections realistic and efficient. This is particularly the case in the US when connections have to be established between states. The router is mandatory to collect the traffic coming from regional networks or national networks.

High throughput routers are required when it is not possible to derive huge pipes from the network. The collected traffic is important and there is no chance to harmonize the traffic at a large scale. It is a model more close to the European model where states are still independent even if they are included in a European community. The growing network will come from states re-exploiting existing infrastructures. So the aggregation between states is less evident and imposes more flexibility to be adapted to a new local strategy. In that case, a backbone based on high throughput routers/switches seems to be the best adapted.

Smart router

Figure C1.3.11 shows the global structure of a smart router. The router collects the traffic at the packet granularity. Packets are put into queues and are sent on a specific wavelength. The optical cross-connect has to manage high throughputs. The structure can be based on a MEM technology. The approach is very interesting when the traffic is strong enough to open large pipes thus enabling the establishment of a waveband. The cross-connection at a waveband level is the best guarantee of simplicity and reliability without creating breakthrough between the transmission system and the switching cross-connect.

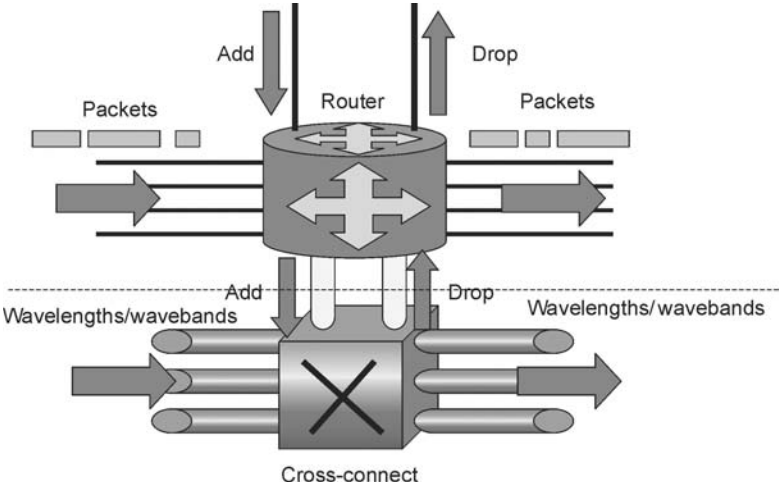
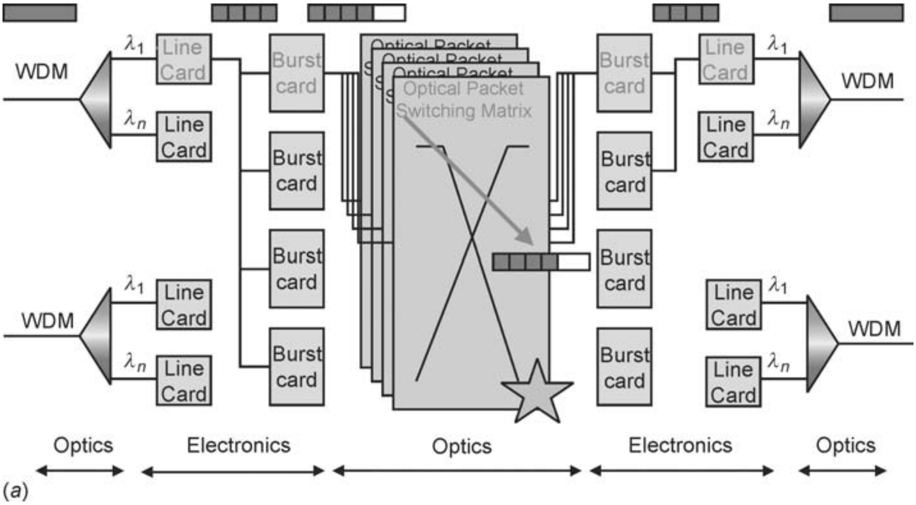


Figure C1.3.11. Smart router schematic.

High throughput routers

Figure C1.3.12(*a*) shows the generic structure of a high throughput router (multi-terabit-class router) exploiting an optical core in its centre part whereas figures C1.3.12(*b*) and (*c*) show a prototype realized and the BER curve. The optical matrix is basically a fast space switch, creating connections at the packet level. The burst card is responsible for the packet format adaptation whereas the line card is used to manage the incoming traffic. Buffers are at the input and at the output of the optical matrix, in the burst cards and in the line cards. An internal speed-up can be exploited to guarantee the full functionality even in case of failure of one of the switching planes. This approach is currently adopted by many constructors.

Both approaches are really important and demonstrate the potential of optical switching for basic functions, mainly focused on space switching: slow or fast.



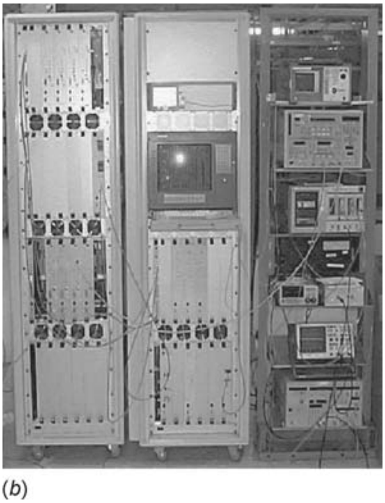
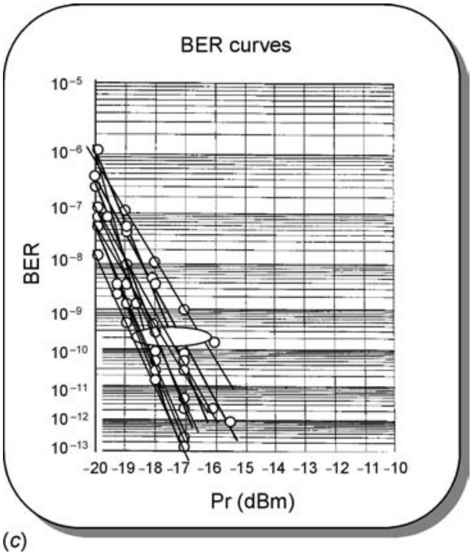


Figure C1.3.12. (*a*) High throughput router schematic. A fast optical switching matrix could be adopted in the centre of the architecture. (*b*) Photo of a 640 Gbit s^{-1} throughput optical matrix. (*c*) BER performance.



(*c*)

Figure C1.3.12. (*Continued*)

Medium term approach: network concept based on O/E/O nodes

In the previous cases, solutions are based on a traffic profile assumption which enables the circuit switching technique for long haul networks or enable the packet technique for small scale backbones. But the problem will occur when the application bit rate is increased together with the variance. This can rapidly happen if merging low bandwidth connections coming from mobile phones and high bandwidth coming from more and more powerful PCs reinforced by an optical connection giving access to very high bit rates per user. In this particular case, the variance can be increased leading to huge problems for efficient aggregation, and forcing the telecom companies to think differently towards an optical packet platform for the backbone.

This means that, due to the traffic profile evolution, the packet will be used extensively. And the key question could be: how to realize an efficient network capable of handling the required capacity while providing a mandatory flexibility?

The generic approach

In the network concept, we mainly exploit the edges to prepare the traffic in such a way that the traffic constraints inside the core of the network are relaxed. This means that all the complex functions are located in the edges like: aggregation, switching per destination, classification, packetization, traffic shaping, load balancing and admission control. The traffic profile, having a better shape, is then sent in the network. The core nodes will be responsible for the synchronization, the contention resolution and the switching. In this case, the packet being created in the edges only simplifies the structure of the core router.

What type of packet format?

As a first introduction, and if possible, it can be envisaged to introduce an existing packet format. The G709 framing is currently being investigated to identify the potential of the concept but other packet formats could be considered.

For a second introduction, it seems clear that a more smaller packet size is required to relax the problems of aggregation. This also imposes a standard on a packet which does not exist today.

It must be noted that some universities and laboratories are currently studying the possibility of managing variable packets called bursts. The advantage of this approach is that the edge part is simplified in its functionality but the core nodes are more complex to control and the overall performance is affected by a highly sporadic traffic profile.

Structure of the core node

The core node is strongly simplified with respect to the first approach since all the complex functions are located in the edge nodes.

Figure C1.3.13 describes a representative structure of such a core router:

It can be noticed that the particularity of this architecture is to have synchronization stages and memory stages before and after the optical matrix in the core of the high throughput routers. This optical matrix has been introduced in the first introduction scenario, so the step is quite easy to cross. The challenge is greater at the management level than at the node level.

Long term solution: all-optical network

In the previous scenario, we still needed a lot of costly O/E conversion. One key question is: can we efficiently reduce the number of O/E conversion stages in the core routers?

For this, we need to solve three key problems:

- The synchronization (to re-align the packets before the switching).
- The regeneration (to enable the cascade of several optical core nodes).
- The contention resolution (to be able to offer the required packet loss rate and delay with respect to the Class of Services requirements).

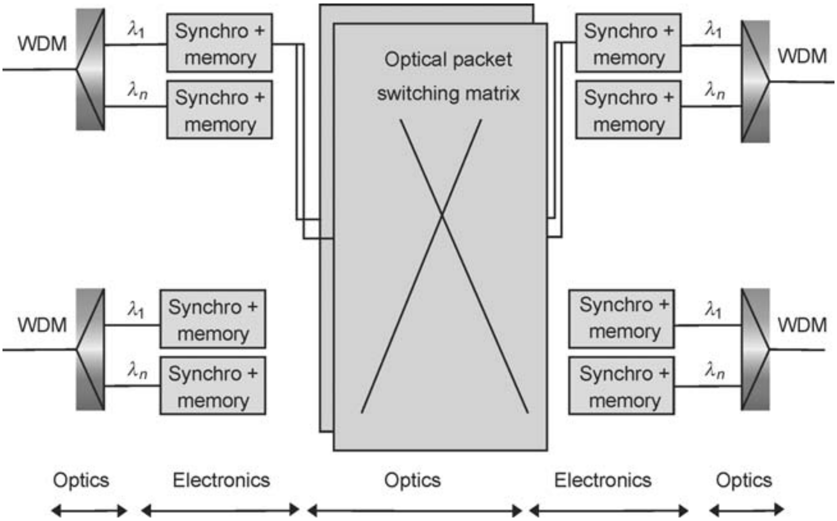

WDM
λ_1
λ_n
Synchro + memory
Synchro + memory
Optical packet switching matrix
Synchro + memory
Synchro + memory
λ_1
λ_n
WDM
WDM
λ_1
λ_n
Synchro + memory
Synchro + memory
Synchro + memory
Synchro + memory
λ_1
λ_n
WDM
Optics
Electronics
Optics
Electronics
Optics

Figure C1.3.13. Schematic of a core router in the concept of a packet network.

Synchronization stage

In electronics we have bit memories making the synchronization process very simple. The information is stored at the distant clock rhythm and it is extracted at the local clock rhythm. However, the structure is quite complex since the process is done at low bit rate imposing thus: one stage of WDM demultiplexing and one stage of bit rate demultiplexing.

As in optics we do not have bit memories and therefore it is important to think differently. By imposing in the packet format a sufficiently large guard band, we simply need to preserve the phase between consecutive packets in order not to have collisions. This means that we need simple synchronization structures capable of having a resolution so as to avoid the problem mentioned. Typically, the resolution that can be handled is in the range of few nanoseconds. It is exactly what we will adopt for the synchronization.

What kind of problem we need to face?

The first problem is the thermal effect in the fibre modifying the index and creating variable delays during the propagation of data depending on the average temperature of the fibre. This means that all the WDM multiplex will be affected. A WDM structure could bring a solution to this problem.

The second problem is that we do not have control of the time jitter created in any optical switching fabric. This packet jitter can be a blocking point in the cascade of several nodes. We need to control the packet jitter packet per packet which indicates that the control must be done at the wavelength level but not at the WDM level.

So in summary, we can easily solve the problem of the synchronization by using one or two stages and combining a processing at the WDM level and a processing at the wavelength level. In both cases we operate at the line bit rate. The gain is in the simplicity of the synchronization process and in the complexity of the structure, making this synchronizer reliable.

Regeneration stage

The regeneration stage is mandatory if we want to exploit the maximum throughput of the optical switching matrix. The regeneration separates the two systems: the transmission and the switching, in order to lead to a maximum throughput for the nodes. It is also the only way to cascade nodes when the line bit rate is high. Once again, as the processing is done at the line bit rate level, the structure of the optical regenerator is really simple, being a guarantee of simplicity and robustness.

Contention resolution

The contention resolution is still an issue in optical architectures since we do not have any efficient optical memory. To solve the problem, we will exploit the WDM dimension and more particularly the statistical multiplexing over the different available wavelengths. So the technique adopted is to avoid collision by re-affecting the wavelength to the packet at the output of the switching fabric. As the number of wavelengths per fibre can be limited, a recirculation buffer is then mandatory to solve the contention properly. By combining both techniques, the performance can easily reach the performance of a classical electronic switch but offering here all the switching capacity in one unique stage.

Optical matrix adapted to optical interfaces

If photodiodes have very low sensitivity it is not the case when introducing all-optical interfaces. So the optical switching fabric must be adapted to these optical interfaces by providing the required power. Once again different techniques can be proposed to achieve this goal.

Generic structure of optical core nodes exploiting optical techniques

A generic structure for an all-optical packet core is described in figure C1.3.14. The particularity of this architecture is that there is no O/E conversion except for the electronic control and the memory, making this architecture cost effective. Based on the previous concept, this architecture is simply an evolution of the core node exploiting optical functions for a better efficiency and a potential line bit rate increases at a lower cost. This approach is fully compatible with future point-to-point transmission systems at 40 Gbit s^{-1}.

C1.3.5.2 Required technology

Available technology

For the short and medium term approaches we need:

- Short term approach: an optical technology for space switching.
    - For optical cross-connects, the MEMs technology is probably the most promising technology.
    - For fast optical matrix, we need to list:
        - For the optical matrix itself: free space-like technologies (Chiaro-like), tuneable source-based technologies (Lucent) and SOA-based technologies (NEC, Alcatel, etc).
        - For the receivers (still in the lab.): packet mode receivers capable of being fully transparent to the packet phase and capable of absorbing packet power variations arriving at the

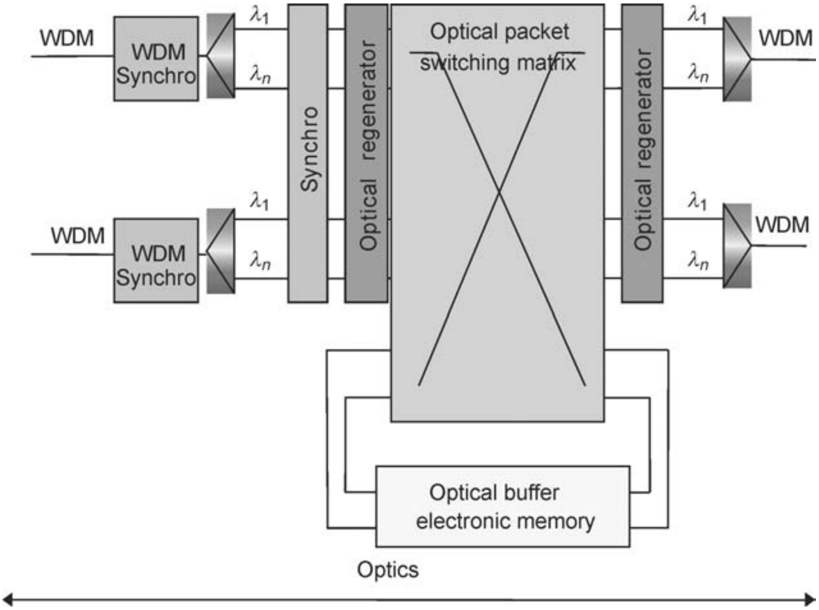

WDM
WDM Synchro
λ_1
λ_n
WDM
WDM Synchro
λ_1
λ_n
Synchro
Optical regenerator
Optical packet switching matrix
Optical regenerator
λ_1
λ_n
WDM
λ_1
λ_n
WDM
Optical buffer electronic memory
Optics

Figure C1.3.14. Structure of an optical core router exploiting the optical resources but including an electronic memory stage in recirculation to guarantee the performance.

packet rhythm. Several companies have proposed such a kind of receiver: Lucent, NTT, NEC, Alcatel, etc.

+ Burst cards (still in the lab): fully electronic adaptation interfaces for the packet format used inside the fabric.

Advanced technology

To realize compact systems, there is first a need for an integrated technology.
To realize the key sub-blocks presented previously we need:

- for the fast optical switching matrix: compact tuneable lasers, SOA gate arrays, integrated wavelength selectors;
- for the optical regenerators: integrated Mach–Zehnder, self-pulsating lasers, etc;
- for optical synchronization: optical gates.

C1.3.5.3 Technological feasibility

Viability of advanced sub-blocks and feasibility issues

Optical matrix

In the case of opto-electronic interfaces, the fast optical switching matrix (introduction scenario for the short and medium term), the constraints are quite relaxed since the sensitivity of the receivers allows the design of switches with small output powers. Typically, reception powers as low as -10 dBm can be considered at the output of the switching fabrics. This also means that the amplification is limited in the core of the switch, leading to very compact and less power consuming architectures.

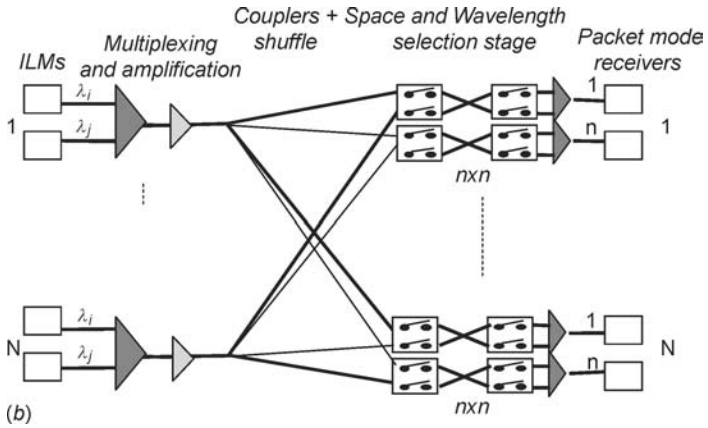
One typical matrix is the SOA-based matrix, requiring simply an amplification stage before the splitting stage is the Broadcast-and-select architecture. Another one is using tuneable lasers and a wavelength router in the centre. Both are represented in figure C1.3.15.

The first architecture (figure C.1.3.15(*a*)) takes advantage of broadcasting functionalities, and exploits robust devices like ILMS or SOAs. However, it is limited in capacity, mainly due to large losses that have to be compensated by amplification. The OSNR affected mainly limits the capacity.

The second architecture (figure C1.3.15(*b*)) has *a priori* a larger potential in terms of capacity since the architecture simply includes a tuneable source and a passive wavelength router. However, this architecture is not adapted to the broadcast of the packets and the fast tuneable laser is probably the most challenging switching element.

In the perspective of the long term scenario, with optical interfaces, the constraint comes from the output power that must be high enough to be compatible with optical interfaces. In addition, the polarization is responsible for problems in the optical regenerative structures, it is then fundamental to transform a switched packet stream into a packet stream in a transmission like configuration. This is the reason why an optical conversion is mandatory in the switching matrix.

Figure C1.3.16 shows an optical matrix based on a SOA technology but including a new element: the wavelength selection/conversion stage, as it is studied in the frame of the IST DAVID project.


Couplers + Space and Wavelength
shuffle
selection stage
Multiplexing
and amplification
ILMs
Packet mode
receivers
λ_i
λ_j
1
1
n
1
nxn
N
1
n
N
nxn
(b)

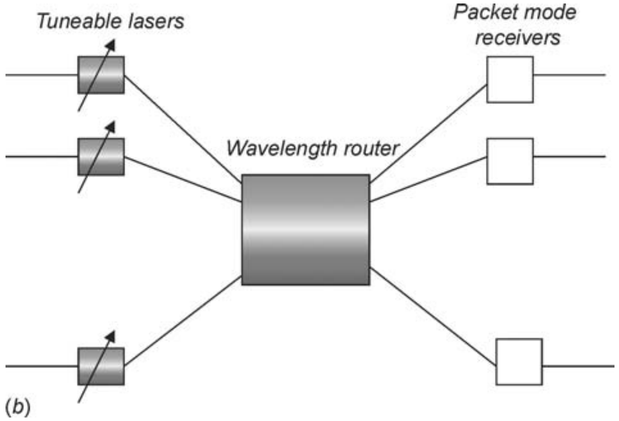

Tuneable lasers
Packet mode
receivers
Wavelength router
(b)

Figure C1.3.15. (*a*) Optical matrix based on SOAs. (*b*) Optical matrix based on tuneable lasers.

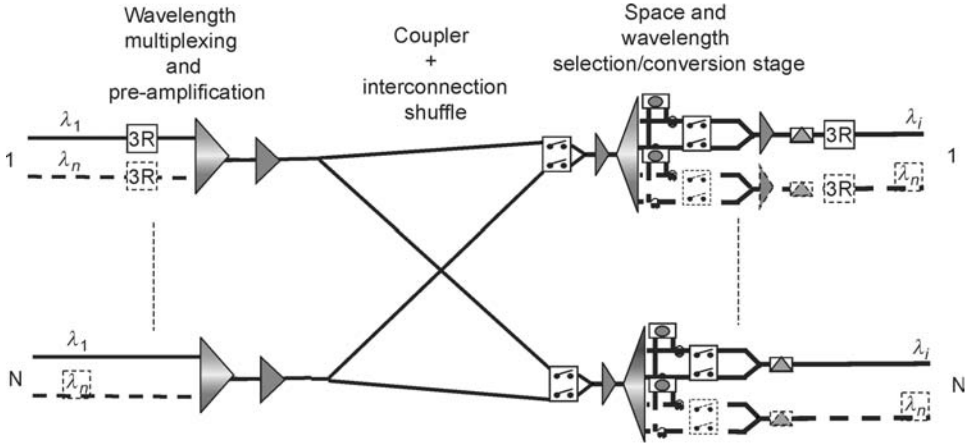

Wavelength
multiplexing
and
pre-amplification
Coupler
+
interconnection
shuffle
Space and
wavelength
selection/conversion stage
λ_1
λ_n
3R
3R
1
λ_i
3R
3R
λ_n
1
λ_1
N
λ_n
λ_i
N
λ_n

Figure C1.3.16. SOA-based matrix compatible for optical interfaces.

SOA technology feasibility

The SOAs have been used in different system applications, for amplification but also for wavelength conversions or for optical gating. To realize large systems as described previously, there is a need for a large amount of components. In this case, the integration is then mandatory to make such a matrix very compact. OPTO+ has designed and realized 32 SOA-gate array modules. The module shown in figure C1.3.17 includes 32 SOAs and their respective drivers. It has been used to realize a 640 Gbit s^{-1} switching matrix.

Tuneable source feasibility

The tuneable source is a key component for many system applications.

In the case of slow switching we can identify the tuneable wavelength conversion to provide the required flexibility to achieve a best utilization of the wavelengths in a network. Another evident application is the replacement of ILMs with tuneable sources. The advantage is mainly in the spare cost: instead of duplicating all the sources, the objective is to have only one source capable of emitting at any wavelength of the comb exploited in the WDM system. Finally, another application is the monitoring of optical switching systems. In this particular case we need a compact structure capable of testing the different wavelengths and paths of a switching system. To be compatible with the system constraints, the requirements are: switching times in the range of milliseconds or more (for monitoring or for sources), large tuneability, high output power, and good extinction ratio.

In the case of fast switching, the main applications are for the metro and the backbone. The tuneability is fundamental in providing the required flexibility to exploit the WDM dimension in optical packet switching network concepts. The requirements are: fast switching time in the range of a few nanoseconds, small tuneability (four or eight channels), high output power, good extinction ratio, and high ON/OFF to guarantee no impact of the crosstalk on the signal quality.

For slow structures, a DBR laser has been tested by different laboratories and feasibility is not an issue.

For fast structures, the main problem is the stability of the wavelength. DBR can be considered if the tuneability is small. These components have been demonstrated to be feasible, with switching times in the range of a few tens of nanoseconds. Another alternative is the selective source. Based on the cascade of a laser array, a SOA gate array, a phasar and one integrated modulator, this structure has been demonstrated to be feasible.

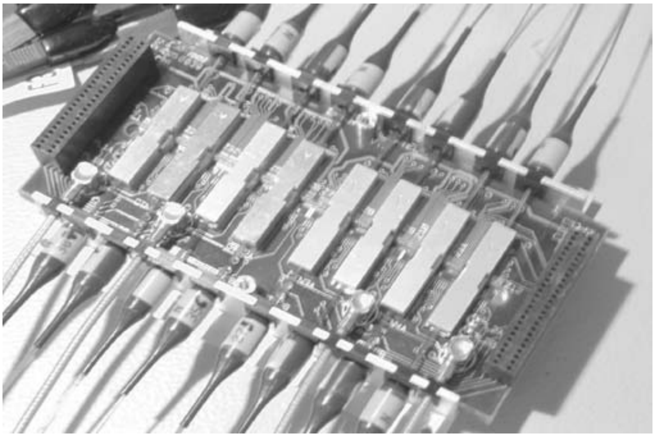


Figure C1.3.17. 32 SOA-gate array module.

Optical synchronization

The optical synchronization is probably the most challenging function. The objective is to process the signal, if possible, in WDM regime or at the wavelength level. The second important point is probably the lack of digital memory which forces designers to think differently. In that context, the synchronization cannot be done at the bit level. We assume that the synchronization can be efficient in a resolution of a few nanoseconds. When we have said that, the other point is to identify the source of de-synchronization with respect to a reference clock.

The first source of loss of synchronization is the thermal effect in the transmission fibre. With a value between 40 and 200 ps km^{-1}, depending on the mechanical protection scheme adopted for the fibre, thermal effects can dramatically affect the phase of the packet streams. The WDM dimension can be advantageously exploited to make the synchronization stage compact and low cost.

The second source is the loss of synchronization in switching fabrics due to a nonideal path equalization. This occurs at the packet level, imposing a synchronization at the wavelength level.

The structure adopted is shown in figure C1.3.18.

Optical regeneration

The optical regeneration is one of the fundamental functions to make the approach realistic. To build all-optical networks while having optical switches capable of handling terabits/second throughputs, the optical regeneration is then mandatory at the periphery of switching architectures. The main functions are the total reshaping of the pulses in the amplitude and in the time domains. To achieve this reshaping, several techniques can be adopted. We will retain one, particularly adapted to the characteristic of switching fabrics creating strong impairments between pules or between group of pulses. The technique adopted is a total reshuffle of the pulses adopting nonlinear elements like Mach–Zehnder or Michelson structures.

The main distortions identified are the following: bits affected at the periphery of packets due to the switching regime, nonlinear effects like cross gain modulation and four wave mixing, crosstalk (in-band and out-of-band), patterning effects by crossing active devices, noise accumulation, jitter accumulation, etc.

To overcome these effects a structure has been proposed in the frame of the DAVID project. This structure, presented in figure C1.3.19, has the following characteristics. By using a cascade of two

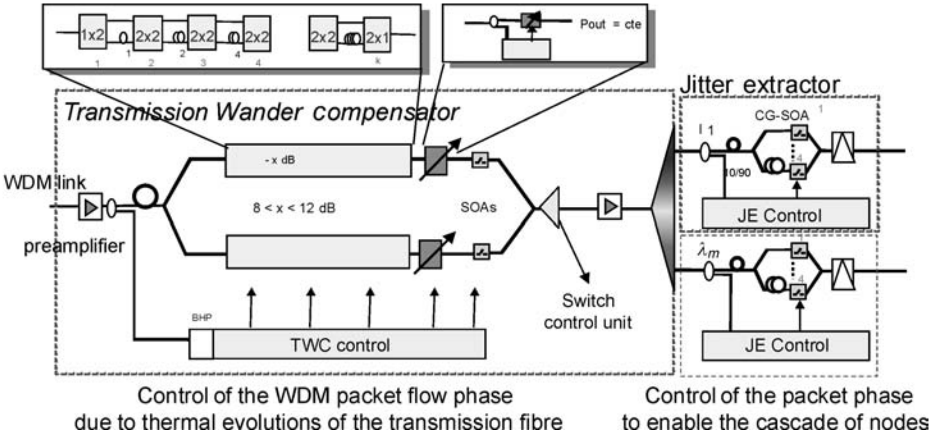

1x2
2x2
2x2
2x2
2x2
2x1
1
1
2
2
3
4
4
k
Pout = cte
Jitter extractor
Transmission Wander compensator
CG-SOA
1
l 1
-x dB
10/90
4
WDM link
8 < x < 12 dB
SOAs
JE Control
preamplifier
λ_m
4
BHP
Switch
control unit
TWC control
JE Control
Control of the WDM packet flow phase
due to thermal evolutions of the transmission fibre
Control of the packet phase
to enable the cascade of nodes

Figure C1.3.18. Optical synchronizer as proposed in the frame of the IST DAVID project.

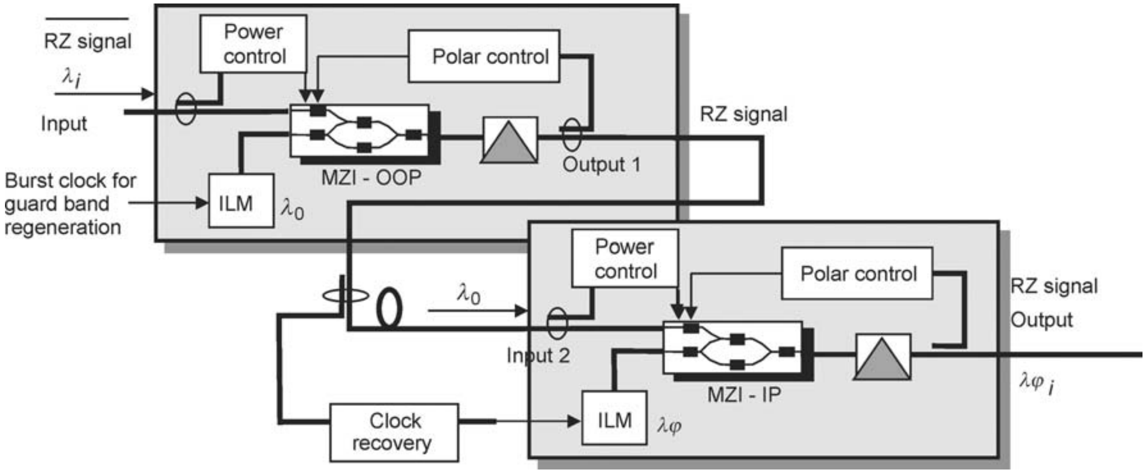

RZ signal
λ_i
Input
Burst clock for guard band regeneration
Power control
Polar control
MZI - OOP
ILM
λ_0
Output 1
RZ signal
λ_0
Input 2
Power control
Polar control
MZI - IP
ILM
$\lambda\varphi$
Clock recovery
RZ signal
Output
$\lambda\varphi_i$

Figure C1.3.19. Structure of the 3R regenerator as it is studied in the frame of the DAVID project.

nonlinear elements, the convoluted function creates a much more nonlinear transfer function thus limiting the noise transferred in the first cascade. This has an important impact on the OSNR specification which can be close to the back-to-back value (before the first stage) even in the case of a large number of cascades.

To really reshuffle the pulses, different techniques will be adopted. We can retain an amplitude and a phase modulation creating an amplitude modulation in interferometric structures to really enhance the extinction ratio and remove the noise. The second technique adopted is a sampling technique of each pulse with a clock to remove the jitter. The wavelength conversion technique will then be preferred to reallocate the wavelength in the correct wavelength comb of the new system.

Feasibility of network concepts

The feasibility of the approach was demonstrated for the first time in 1998, at the end of the ACTS KEOPS project.

In this project, we have cascaded 40 network sessions error free, at 10 Gbit s^{-1} per wavelength demonstrating for the first time the possibility of building an all-optical network at a backbone scale.

Figure C1.3.20 gives the network session tested and put in a loop to demonstrate the concept.

C1.3.5.4 Performance expected

The performance is probably one of the most important indicators for the feasibility of such concepts. When the physical aspects are verified, the challenge becomes the performance in a real traffic environment.

Environment and specificity of the backbone

If, in the metro, the capacity is limited, in the backbone this is the major characteristic. To provide the capacity with a technology limited today to 10 Gbit s^{-1}, the only solution is in the exploitation of the WDM dimension.

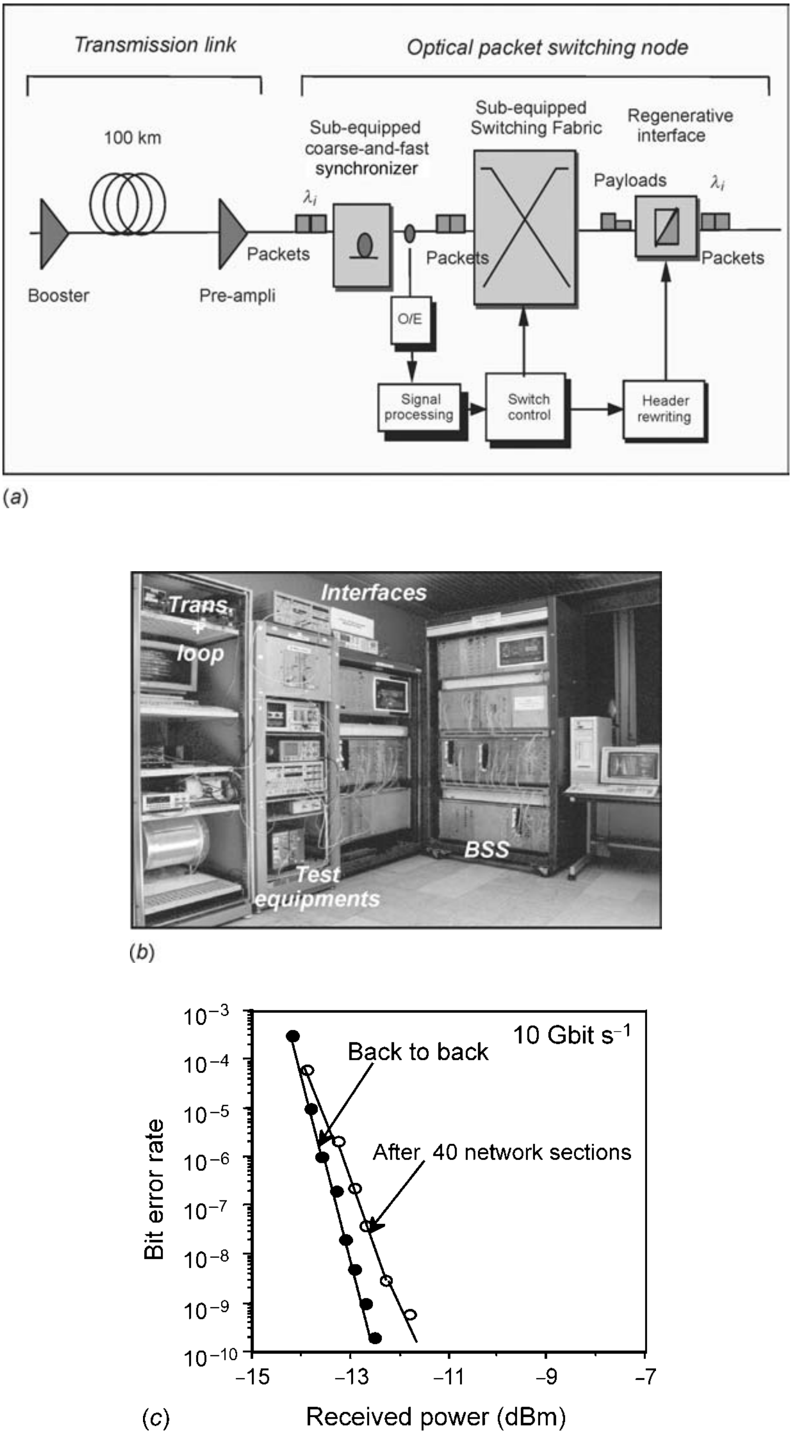

Transmission link
Optical packet switching node
100 km
Booster
Pre-ampli
Packets
Sub-equipped coarse-and-fast synchronizer
λ_i
Packets
O/E
Signal processing
Sub-equipped Switching Fabric
Switch control
Regenerative interface
Payloads
λ_i
Packets
Header rewriting
(a)
Trans. + loop
Interfaces
BSS
Test equipments
(b)
10 Gbit s^{-1}
Back to back
After 40 network sections
Bit error rate
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
−15
−13
−11
−9
−7
Received power (dBm)
(c)

Figure C1.3.20. (*a*) Network session put in a loop to test the feasibility of an all-optical network. (*b*) Photo of the demonstrator. (*c*) BER curves giving the physical performance.

So the WDM dimension will be fully exploited to provide the required capacity but also to avoid collisions due to the natural statistical multiplexing of packets on the wavelengths.

The second particularity is the aggregation. Depending on the traffic profile, circuit switching or packet switching will be preferred.

Circuit switching techniques for an immediate introduction

Circuit switching techniques can be envisaged in a first scenario as a transport layer to provide the capacity of transport.

To be compatible with a DATA traffic, the coupling of a cross-connect with a packet router is, even today, the more pragmatic approach. This is a subject of strong interest for products that are used at present.

However, this solution is not really cost effective because only two alternatives can be adopted:

- All the wavelengths are connected to the packet router, and in this case the number of TX/RX dramatically increases the cost per port.
- Only a part of the wavelengths is connected to the packet router and in this case the traffic matrix needs to be very stable. The deterministic approach for the number of connections becomes nonrealistic when the traffic profile becomes highly statistical.

Optical packet switching techniques to cope with a traffic profile evolution

In this case what could be the benefit of the concept proposed?:

- The packetization at the edge level with an optimized size can reduce the latency in the creation of the packets. A second technique to accelerate the filling ratio is in the upgrade of the best effort in a premium class of service. So the advantage of this solution is that latency can be controlled to reduce the latency in the rest of the network. Calculations show that there is a global benefit in terms of end-to-end latency.
- The exploitation of the WDM dimension once again reduces the latency. The packets cross the architecture and they see only a transmission path, even if it is switched. No buffers are crossed so the resultant latency is minimum.

Figure C1.3.21 shows a table summarizing the performance in terms of packet loss rate established in the frame of the ROM project. It appears that on the three class of services considered, the end-to-end

Total load	CoS 1 PLR	CoS 2 PLR	BE no recir. PLR	BE Recirc 16λ ,PLR
0.5	<10e-6	<10e-6	~10e-04	<10e-06
0.7	<10e-6	<10e-6	0.012	<10e-06
0.8	<10e-6	<10e-6	0.03	<10e-04
0.9	<10e-6	~10e-5	0.065	0.007

Figure C1.3.21. Performance of the all-optical packet switching network concept.

performance can be obtained. From dimensioning issues it appears that for the WAN a sum of 30% for CoS 1 and CoS 2, and a BE lower than 80% is tolerated.

This demonstrates the viability of an all-optical concept and as a consequence the viability of the opto-electronic scenario.

C1.3.6 Introduction on the market: criteria

C1.3.6.1 Criteria of selection for a new technology

To select a technology to task is not easy but we can draw some conclusions:

- Bit rate evolution at the user part creating a convergence of the bit rate in all the layers of the network and forcing a transfer of the traffic profile even in the backbone. This will create a need for high flexible networks to cope with a traffic profile and not with a traffic matrix. Packet technique is today the only pragmatic solution with a co-existence of circuit switching techniques.
- The key bit rate is 10 Gbit s^{-1}. All the companies are focusing on 10 Gbit s^{-1} which will develop a volume to make this technology compact and cost effective. This also reinforces the packet technique, because the bit rate is now totally independent of the physical bit rate; the granularity is offered by the packet size and not by the bit rate of the wavelength.
- If a circuit switching technique is adopted today, it must be compatible with a migration towards packet switching.

C1.3.6.2 Cost approach

For the cost approach, everything will depend on the aggregation efficiency. In the following we have computed the relative cost of different approaches, comparing mainly packet switching and circuit switching.

Metro part

If the average load of a wavelength is high enough, due to an efficient aggregation process, then circuit switching is probably viable. But if the load is low, below 20%, even if the cost of switches are more expensive, the gain in statistical multiplexing creates a real opportunity for packet techniques making them less expensive than circuit switching techniques.

The main reason for this gain is probably the high cost of the wavelength due to expensive infrastructure costs, pushing all telecom companies to prefer an increase of the bit rate rather than an exploitation of the WDM dimension.

So the tendency is probably packet techniques to decorelate the bit rate from the granularity of switching, and high bit rates to adopt the most cheap technology while providing the required capacity.

Figure C1.3.22 shows the areas where optical packet switching is better than circuit switching.

The number: 2, 4, 6 and 8 indicates the ratio in terms of cost per port (wavelength) between an optical packet switch and a cross-connect targeting the same size (256 × 256) and the same technology.

The load of a wavelength is the average load.

The ratio on the horizontal axis is a ratio between the wavelength transmission cost (including the installation costs) and the cross-connect port.

For example, if the ratio between the cost of a wavelength in the transmission system and the cost of a cross-connect port is equal to 1 (red bar), optical packet switching techniques are interesting:

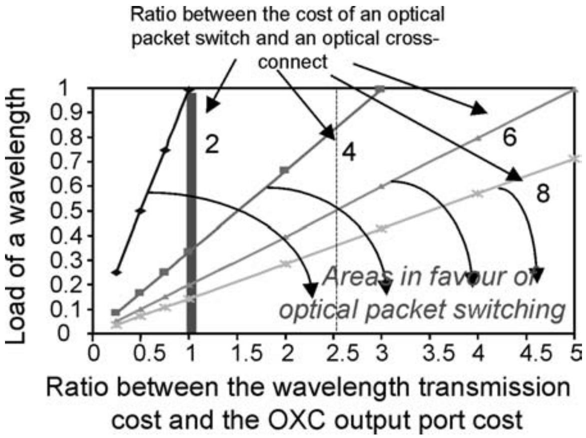


Figure C1.3.22. Need to exploit packet techniques in the near future.

- 2: always, whatever the load is;
- 4: if the average load of a wavelength in circuit switching is lower than 33%;
- 6: if the average load of a wavelength in circuit switching is lower than 20%;
- 8: if the average load of a wavelength in circuit switching is lower than 13%.

So the tendency is the following: if the cost of a wavelength in a transmission system is high (the case of the metro where the installation cost is not negligible), or if the aggregation is not efficient enough forcing an average load very low (this is a serious tendency with the increase of the application bit rate, and the sporadicity of the traffic profile), packet switching techniques always exhibit a better performance than circuit switching.

Backbone part

As an example we have computed, for two levels of aggregation, the load of a network with respect to the distance of the network (for the access to the backbone). It appears that in major cases, if the aggregation process is not enough, even in the backbone, the packet switching technique is the cost effective solution.

Figure C1.3.23 shows the importance of packet switching techniques, also for the backbone. This is one of several curves that could be drawn. However, it once again shows a tendency.

The grooming or the aggregation efficiency depends on the traffic profile in large part. So we plot two indicative curves:

- One exhibiting an efficient aggregation (a realistic case is when the CBR is higher than VBR or when the number of connection points is high to facilitate the grooming process);
- One exhibits a less efficient aggregation (a realistic case is when the VBR becomes dominant). In that case we cannot have a stable traffic matrix, and we are addressing a sporadic traffic profile (a realistic case if we have a bit rate convergence from the access to the backbone).

The vertical axis indicates that the required average load of a wavelength is cost effective.

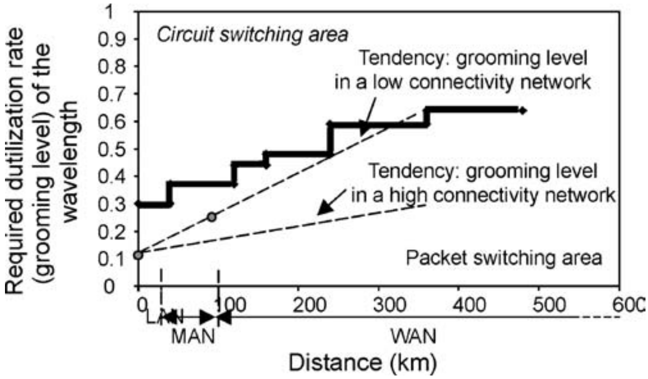


Figure C1.3.23. Curve showing strong interest to introduce packet switching techniques even in the backbone.

The horizontal axis indicates the average distance of a transmission system with respect to an average network session representative of the network considered. So the WAN starts for transmissions higher than 100 km. The calculations show the importance of the time multiplexing. If the distance is long, there is a large number of clients sharing the same network infrastructure. So the cost per client is reduced. In addition, the cost of the installation of a wavelength is considered lower than for the metro. The reason is because in the WAN, natural infrastructures are exploited to reduce the installation costs (like highways, or railways, etc). If we are under the curve in bold, there is an advantage to introducing packet switching techniques. The grooming tendency gives values for the required load in circuit switching.

For example, in the case of a good grooming efficiency, if the average distance of propagation of a representative network session is lower than 300 km, you will have a cost gain by exploiting packet techniques. In the case of a low grooming efficiency, packet switching techniques are always more efficient.

C1.3.7 Conclusions

In this chapter, we have suggested the optical switching technique as a potential technique for the next generation of systems or networks. But, more importantly an evolution scenario is given for the metro part and the backbone part describing what could be the most promising solutions. Optical packet switching techniques appear very attractive since they really offer a solution compliant with the traffic constraints.

Circuit switching techniques will be introduced as a first step, but we must not forget optical packet switching techniques that will improve the bandwidth utilization.

We have seen that there is no problem building any of the network concepts proposed, since all the functions have already been demonstrated to be feasible. The solution is now in the availability of the technology and in the cost. The progress on this integrated/low cost optical technology will be fundamental for the future systems and could really provide new advantages with respect to classical solutions exploiting electronics only.

Today we can imagine two scenarios:

The first one will consist of the introduction of a circuit switching platform to give a concrete answer to an immediate need at a lower cost. Circuit switching is probably the best today. However, we cannot forget the evolution of the traffic profile to increase the bit rate at the access part. So the migration scenario is an important argument to propose solutions that can be rapidly adapted to packet switching techniques with the best flexibility and upgradability.

The second scenario is in the adoption of packet switching techniques like RPR for the metro or routers for the backbone. And we then need to think about competitive solutions with serious added values to justify the introduction of optical techniques in the network. Optical packet switching is probably one technique that can emerge. In the metro part, the benefit is mainly in the exploitation of the WDM dimension and in the very simple in-line processing (without any buffer) to reduce the latency and the number of TRX. In the backbone part, the benefit is probably in the adoption of large packets assimilated to containers in order to be able to exploit techniques to reshape and manage the traffic profile in the edge nodes, and WDM techniques to reduce mainly the latency without constraining the capacity expansion in the core nodes.

But to build sub-systems there is also a need for an advanced technology. Without any advanced technology like tuneable source or tuneable filters there will be no chance to provide the functionality required to be really competitive on other aspects. So the development of this new technology (components and systems) is then fundamental and will position a constructor of equipment as a leader in the future market.

Acknowledgments

The author acknowledges colleagues from Alcatel, the European Commission and the French ministry for funding for the following projects: RACE 2039 ATMOS, ACTS 043 KEOPS, REPEAT, IST DAVID, RNRT ROM, particularly T Atmaca from INT, M Renaud from Opto + who provided key results in terms of network performance and optical component illustration and all the partners involved in these projects.

Further reading

Brackett C A 1996 Is there an emerging consensus on WDM networking *J. Light Technol.* **14** 936–941

Gambini P 1997 State of the art of photonic packet switched networks *Photonic Networks* ed G Prati (London: Springer) pp 275–284

Blumenthal D J *et al* 1994 Photonic packet switches: architectures and experimental implementations *Proc. IEEE* **82** 1650–1667

Misawa A *et al* 1996 40 Gbit/s broadcast-and-select photonic ATM switch prototype with FDM output buffers *Proc. ECOC'96* **4**

Renaud M *et al* 1997 Network and system concepts for transparent optical packet switching *IEEE Commun. Mag.* **35** 96–102

Guillemot C *et al* Transparent optical packet switching: the European ACTS KEOPS project approach *Special issue of J. Light Technol.* at press

Callegati F 1997 Which packet length for a transparent optical network? *SPIE Symposium on Broadband Networking Technologies* (Dallas, USA, November 1997)

Callegati F 1997 Efficiency of a novel transport format for transparent optical switching *IEEE ICT 97* (Melbourne, Australia, April 1997)

Bostica B *et al* 1997 Synchronization issues in optical packet switched networks *Photonic Networks* ed G Prati (London: Springer) pp 362–376

Hunziker *et al* 1995 Self-aligned flip chip packaging of tilted semiconductor optical amplifier arrays on Si motherboard *Electron. Lett.* **31** 488–490

Janz C *et al* 1998 Low-penalty 10 Gbit/s operation of polarization-insensitive Mach–Zehnder wavelength converters based on bulk-tensile active material *OFC'98* WB1 (San Jose, California, Feb. 1998) pp 101–102

Danielsen S L *et al* 1996 Bit error rate assessment of 40 Gbit/s all-optical polarization independent wavelength converter *Electron. Lett.* **32** 1688–1689

Mestric R *et al* 1997 Up to 16 channel phased array wavelength demultiplexer on InP with -20 dB crosstalk *ECIO '97* paper EThE3 (Stockholm, Sweden, 1997)

Zucchelli L *et al* 1996 New solutions for optical packet delineation and synchronization in optical packet switched networks *ECOC'96* vol 3 (Oslo, Norway, Sept. 15–19, 1996) pp 301–304

Chiaroni D *et al* 1997 Feasibility assessment of a synchronization interface for photonic packet-switching systems *ECOC'97* (Edinburgh, UK, Sept. 1997)

Guillemot C *et al* 1995 A two stage transparent packet switch architecture based on wavelength conversion *ECOC'95* vol 2 (Brussels, Belgium, Sept. 17–21, 1995) pp 765–768

Hansen P B *et al* 1997 20 Gbit/s experimental demonstration of an all-optical WDM packet switch *ECOC'97* vol 4 (Edinburgh, UK, Sept. 1997) pp 13–16

Gabriagues J M *et al* 1995 Performance evaluation of a new photonic ATM switching architecture based on WDM *Australian Telecommunication and Network Application Conference* (Sydney, Australia, Dec. 1995)

Chiaroni D *et al* A 160 Gbit/s throughput photonic switch for fast packet switching systems *Proc. Photonics in Switching'97* paper PWB3 (Stockholm, Sweden, April 2–4) pp 37–40

Chiaroni D *et al* 1997 Demonstration of full optical regeneration based on semiconductor optical amplifiers for large scale WDM networks *postdeadline ECOC'97* (Edinburgh, UK, Sept. 1997)

C2.1
Camera technology

Kenkicki Tanioka, Takao Ando and Masayuki Sugawara

C2.1.1 The camera tube and camera

C2.1.1.1 Introduction

The history of photoconductive camera tubes using the internal photoelectric effect began in 1950 with the Vidicon camera tube [1]. A photoconductive camera tube performs both photoelectric conversion and signal storage on a photoconductive target, which is a vapour-deposited film of Sb_2S_3 in Vidicon tubes. Although the Vidicon tube boasts a simple, small and lightweight structure compared with the Image Orthicon tube that used the external photoelectric effect, it also suffers from several weak points such as large lag and dark current. The Vidicon, as a result, has not found much use in broadcasting-type television cameras that require high levels of picture quality. These weak points stem from the fact that excited carriers are easily trapped in Sb_2S_3 deposited film and that charge is injected into the target from an external electrode (injection-type target). On the other hand, the Plumbicon camera tube announced in 1963 features a target formed by a PbO film with a p–i–n structure that blocks the injection of charge from an external electrode (blocking-type target) [2]. With this type of target, the Plumbicon became the first photoconductive camera tube to feature low lag and low dark current among other superior features. In the 1970s, the Plumbicon rode the wave of change to colour television broadcast facilities and became the leading tube for broadcast-class colour television cameras replacing the Image Orthicon tube that used the external photoelectric effect. The research and development of photoconductive camera tubes had been quite active. The 1970s, for example, saw the back-to-back development and commercialization of various blocking-type photoconductive camera tubes including the Chalnicon using CdSe, $CdSeO_3$ and As_2S_3 as targets [3], the Saticon using Se–As–Te [4, 5].

In the 1990s, the solid-state imaging device such as the CCD has become the mainstream image sensor even for cameras used in the field of HDTV broadcasting. There is still a demand, though, for the camera tube, which was once the predominant type of image sensor in the form of the Plumbicon, Saticon, etc for purposes of camera maintenance. As a consequence, modern camera tubes normally employ past technology. An exception, however, is the high-gain avalanche rushing amorphous photoconductor (HARP) camera tube developed by the NHK and Hitachi, Ltd whose novel technology has become the focus of attention [6]. This camera tube achieves a level of sensitivity higher than that of CCDs and conventional pickup tubes by converting light to an electric signal in a photoconductive target and simultaneously amplifying that signal by an avalanche multiplication effect. It features higher quality pictures than past ultrahigh-sensitivity image sensors using image intensifiers. The HARP camera tube can be used in high-definition cameras and for a wide range of applications including the capturing of astronomical phenomena such as auroras and solar eclipses.

The following section describes the mechanism of the photoconductive camera tube covering past types and the ultrahigh-sensitivity HARP tube based on a new operating principle.

C2.1.1.2 Basic configuration and operating principle of camera tubes

Camera tubes come in two main types: an image camera tube that uses an external photoelectric (photoemission) effect and a photoconductive camera tube that uses the photoconducting effect, a type of internal photoelectric effect. For the image camera tube, a typical example is the Image Orthicon developed during the monochrome television age. For photoconductive pickup tubes, there is the Vidicon, which can be called the original tube of this type, and tubes like the Plumbicon and Saticon that played a great role in improving the performance of broadcast colour cameras and achieving hand-held video cameras. The HARP camera tube to be described here is also a photoconductive pickup tube.

The basic configuration of the photoconductive camera tube is shown in figure C2.1.1. The tube consists of a photoconductive target that performs photoelectric conversion and charge storage, and a scanning electron beam system for reading stored charge. The operation of this tube is described later.

Referring to figure C2.1.2, light incident on the target generates electron–hole pairs in the film (using, for example, a blocking-type target as described later). Here, for an ordinary pickup tube, the scanning electron beam irradiates the target at low velocity, and voltage is applied in such a way that

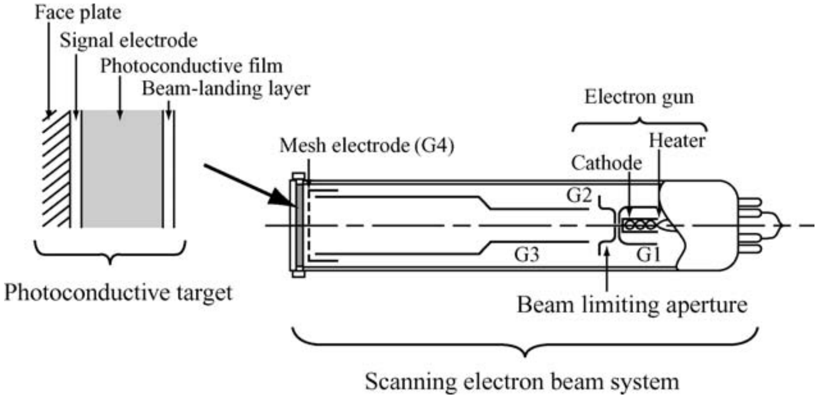

Face plate
Signal electrode
Photoconductive film
Beam-landing layer
Photoconductive target
Electron gun
Heater
Mesh electrode (G4)
Cathode
G2
G3
G1
Beam limiting aperture
Scanning electron beam system

Figure C2.1.1. Basic structure of photoconductive camera tube.

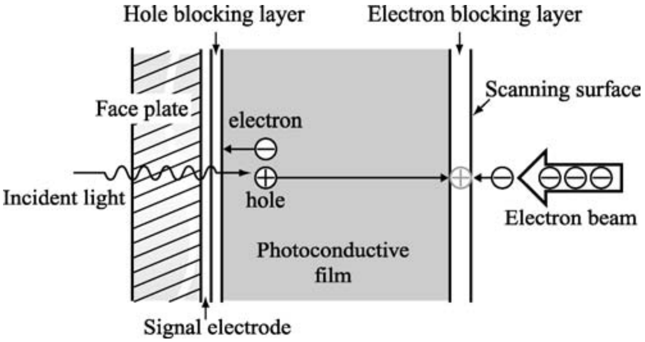

Hole blocking layer
Electron blocking layer
Scanning surface
Face plate
electron
Incident light
hole
Electron beam
Photoconductive film
Signal electrode

Figure C2.1.2. Behaviour of charge in target.

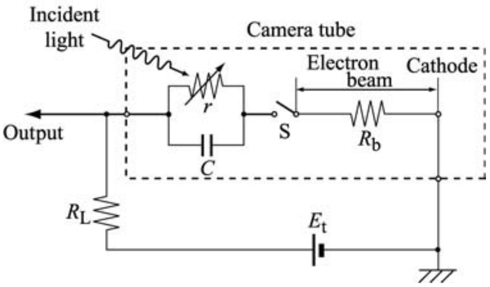


Figure C2.1.3. Equivalent circuit for one pixel in camera tube.

the signal-electrode side takes on a positive potential with respect to the scanning surface. This causes electrons to move towards the transparent signal electrode while holes move towards the target's scanning surface. The material used for the target's photoconductive film, however, generally has a high value of resistance with resistivity at $10^{12}\,\Omega\,\text{cm}$, and these charges accumulate at both ends of the target as a result. Meanwhile, the scanning electron beam system works to focus and deflect the electron beam emitted from the electron gun so as to make it incident on the target at low velocity (low-velocity beam landing). This causes stored holes to recombine and disappear and an equivalent amount of signal current to flow from the transparent signal electrode.

We can make this operation even easier to understand by focusing on a single pixel and using the equivalent circuit shown in figure C2.1.3 consisting of resistors, a capacitor and other elements. In the figure, the symbols r and C correspond to the resistor and capacitor making up the pixel. Here, the value of resistor r changes according to the intensity of the incident light. In addition, R_b denotes the equivalent resistance of the scanning electron beam, S the scanning switch, R_L the load resistance provided externally and E_t the power supply for applying voltage to the target. The arrival of the scanning electron beam on a certain pixel corresponds to the closing of switch S at which time C charges via r. Conversely, departure of the beam from that pixel corresponds to the opening of switch S at which time the charge accumulated in C discharges via r. The discharge period is determined by the number of frames per second and is 1/30 s (1 frame's worth) in principle in the NTSC system. In actual camera tubes, however, the beam is broad compared to the scanning line interval, and beam scanning overlaps the odd and even fields. As a consequence, the discharge period becomes 1/60 s, or one field's worth, despite interlaced scanning. Because the value of r changes according to the intensity of incident light, the amount of discharge is large for a bright subject and small for a dark one. This means that the charging current flowing to C from the power supply when S is closed is equivalent to the current, i.e. signal current, modulated by the brightness of the subject.

Types of scanning electron beam systems and their features

The scanning electron beam system of the camera tube described here is classified in terms of electric-field/magnetic-field combinations used for focusing and deflecting the beam. There are the electromagnetic-focusing/electromagnetic-deflection (MM) type, the electrostatic-focusing/electromagnetic-deflection (SM) type, the electromagnetic-focusing/electrostatic-deflection (MS) type, and the electrostatic-focusing/electrostatic-deflection (SS) type as shown in figure C2.1.4, respectively. Of these, the MS type features high resolution up to the corners of the screen in principle. The SS type, moreover, requires no coil for focusing and deflection and can therefore achieve a compact, light and low-power configuration.

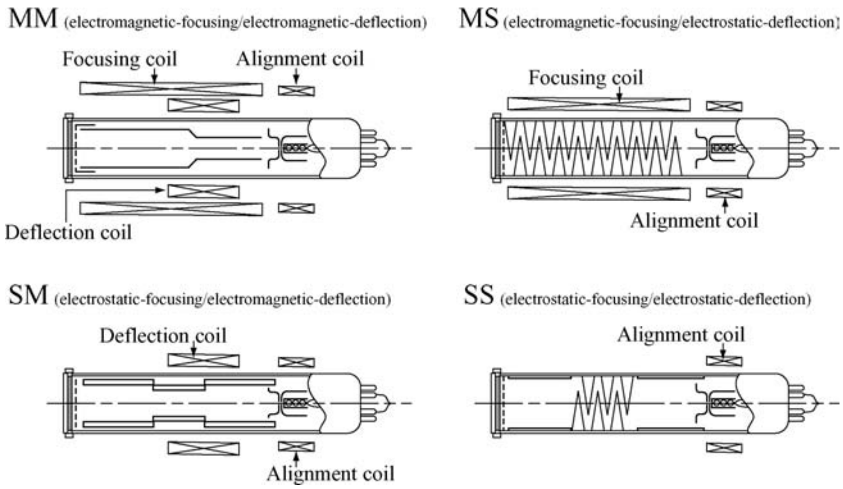

MM (electromagnetic-focusing/electromagnetic-deflection)
Focusing coil
Alignment coil
Deflection coil
MS (electromagnetic-focusing/electrostatic-deflection)
Focusing coil
Alignment coil
SM (electrostatic-focusing/electromagnetic-deflection)
Deflection coil
Alignment coil
SS (electrostatic-focusing/electrostatic-deflection)
Alignment coil

Figure C2.1.4. Focusing and deflection system of camera tube.

Injection and blocking types of targets

Targets can be divided into injection type and blocking type. The structure of the injection-type target is such that charge comes to be injected into the photoconductive film from both the signal-electrode side and electron-beam-scanning side or from either one of these sides. As a result, an amplification effect called 'injection amplification' occurs within the target and high sensitivity with a quantum efficiency of 1 or greater can be obtained. (Quantum efficiency is defined here as the number of output electrons per number of unit incident photons in the target; it is denoted as η.) This injection amplification effect is described here using the target shown in figure C2.1.4 in which electrons are injected from the beam scanning side. Now, one hole created by one incident photon will come to be stored on the beam scanning side as shown in the figure. Then, when the scanning electron beam comes to read this hole, the hole will not immediately recombine with an electron but will instead have to wait until N electrons are first injected into the target until it can recombine with the $(N+1)$th electron. This means that $N+1$ electrons flow out of the target from the signal electrode. In other words, this operation provides an amplification effect with a gain of $N+1$ in which electrons greater than the number of incident photons can be read out to an external circuit. Before the invention of the HARP target, the injection-type target was researched as the only photoconductive target that could achieve high sensitivity of $\eta > 1$. This target, however, suffers from sharp increases in lag and dark currents under high-sensitivity operation and consequent drops in picture quality, and could not, as a result, be viewed as a new approach to camera tubes (figure C2.1.5).

On the other hand, a blocking-type target has a structure in which both the signal-electrode side and electron-beam-scanning side block the injection of charge from the outside. As an example, figure C2.1.6 shows a Saticon target whose main component is amorphous selenium (a-Se). Here, the injection of holes is blocked at the junction between the 1st layer consisting of Se + As (arsenic) film and the SnO_2 (tin oxide) signal electrode and CeO_2 (cerium oxide) film. The injection of electrons, meanwhile, is blocked by the 5th layer consisting of Sb_2S_3 (antimony trisulfide) on the beam scanning side. In addition, the 2nd layer in the figure is an intensifying layer; layer 3 plays the role of conducting holes created in the 1st and 2nd layers to the 4th layer; and the 4th layer acts to decrease storage capacitance in the target and reduce capacitive lag. A target such as this that blocks the injection of charge from external electrodes means that increase in dark current will be small even for an increase in applied voltage. A sufficient

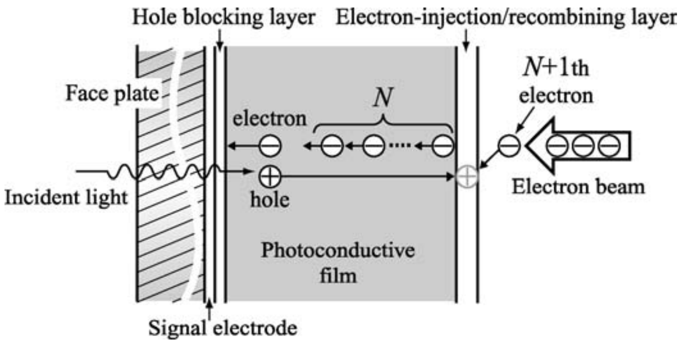

Hole blocking layer
Electron-injection/recombining layer
Face plate
N+1th electron
N
electron
Incident light
hole
Electron beam
Photoconductive film
Signal electrode

Figure C2.1.5. Operating principle of electron-injection–amplification target.

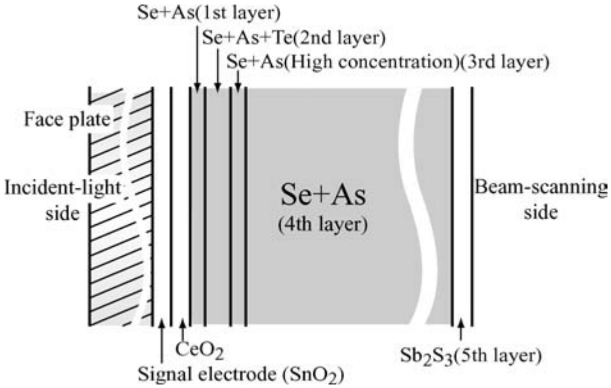

Se+As(1st layer)
Se+As+Te(2nd layer)
Se+As(High concentration)(3rd layer)
Face plate
Incident-light side
Se+As (4th layer)
Beam-scanning side
CeO2
Signal electrode (SnO2)
Sb2S3(5th layer)

Figure C2.1.6. Structure of Saticon target.

electric field ($1.25 \times 10^7\,\mathrm{V\,m^{-1}}$ in the Saticon) can therefore be given to the photoconductive film. As a result, most electron–hole pairs excited by incident light can be separated by a strong electric field to form a signal current, and a relatively high level of sensitivity can be obtained as a consequence. This kind of target also features little dark current and low lag. For the above reasons, targets for recently developed photoconductive camera tubes have been of the blocking type for which high picture quality can be obtained. In a target of this type, however, a scanning-beam electron immediately recombines with a hole stored on the scanning side and subsequently disappears, as shown in figure C2.1.2. To put it another way, the number of scanning-beam electrons landing on the target per one hole is simply one, which in turn means that no more electrons than the number of incident photons can, in principle, be read to the outside, i.e. the limit of sensitivity is $\eta = 1$. In the past, this was referred to as the sensitivity barrier in blocking-type targets.

C2.1.1.3 Ultrahigh sensitivity photoconductive camera tube

The more sensitive imaging devices are, the better they are able to produce clear pictures even in low lighting conditions. Consequently, achieving increased sensitivity has always been an important theme throughout the 1970s or so years of research into imaging devices, and even today it is a matter of fierce competition between researchers.

From the 1960s to the 1980s, NHK (Japan Broadcasting Corporation) researched and developed a variety of high-sensitivity imaging devices such as secondary electron conduction (SEC) tubes and I-CCDs, which are made by combining image intensifiers with charge-coupled devices (CCDs).

But since these conventional high sensitivity devices had problems associated with their picture quality, such as high noise levels and poor resolution, demand grew in the 1980s for the development of imaging devices that combine high sensitivity with high picture quality. During this period, HDTV camera using camera tubes such as Saticon began to be used in practical applications. However, their sensitivity was still rather poor, and as reporting breaking news programs and science programmes began to attach increasing importance to camera sensitivity even for standard TV broadcasts, it became even more important to develop a TV camera with high sensitivity and high picture quality, capable of producing clear images even from poorly lit subjects.

Consequently, from about 1980, NHK began a fresh study with the aim of realizing a high sensitivity and high picture quality imaging device suitable for use in HDTV applications. This study focused on using the amplification effect of an a-Se to obtain a high level of sensitivity.

In 1985, it was found that when an a-Se target in the camera tube is operated in a strong electric field of about $10^8\,\mathrm{V\,m^{-1}}$, continuous and stable avalanche amplification takes place, allowing high sensitivity to be obtained with little picture degradation. Based on this discovery, NHK and Hitachi, Ltd went on to develop a new kind of imaging device called HARP camera tube, which has been studied to this day to achieve further increases in sensitivity and a wider range of applications.

HARP camera tubes, which have achieved sensitivities roughly 100 times greater than CCDs, are used not only for standard TV broadcasts, but also in HDTV hand-held cameras, and are used in the production of night-time news flashes and special programmes such as imaging the aurora.

The following section summarizes the research conducted so far into HARP camera tubes, and describes the features of ultrahigh sensitivity cameras that use them.

C2.1.1.4 The development of HARP camera tubes

The investigation of high sensitivity and high picture quality imaging devices

The basic technique employed in conventional high sensitivity imaging devices involves directing the incident light towards a photocathode and accelerating the photoelectrons emitted from this surface with a large voltage inside a vacuum. For example, in the I-CCD mentioned earlier, these accelerated electrons impinge on a fluorescent surface where they form a bright picture that is imaged using a CCD. Using a photocathode has the advantage that a high level of sensitivity can be obtained quite easily, so it has also been developed in various other types of high sensitivity imaging devices, such as the silicon intensifier target (SIT) tube [7].

But conventional high sensitivity imaging devices that use a photocathode also suffer from drawbacks, such as the following:

(i) Since they use the external photoelectric effect for photoelectric conversion, it is difficult to increase their conversion efficiency to values close to 100%. A low conversion efficiency results in increased picture quality degradation due to shot noise.

(ii) Picture quality can also be degraded by other forms of noise that are characteristic to the device, such as ion feedback noise that arises from residual gases inside the tube.

(iii) It is difficult to achieve the high resolution needed for HDTV cameras with a compact imaging device.

To address these problems, NHK decided to work on developing high sensitivity and high image quality imaging devices that do not rely on the use of a photocathode, and to investigate the possibility of achieving substantial increases in the sensitivity of the photoconductive target in Saticon tubes that were also used in HDTV cameras.

In the mid-1980s when this investigation got underway, the mainstream of imaging devices had started to shift from the camera tubes to solid-state devices (CCDs). Not only are CCDs compact, lightweight, easy to use and highly reliable, but it is also possible to suppress the noise from their internal amplifier circuits to a much lower level than can be achieved with the external amplifiers used with camera tubes. CCDs therefore seemed to have greater potential than camera tubes in terms of sensitivity.

Nevertheless, NHK decided to take a fresh look at photoconductive camera tubes because it was considered the targets in these camera tubes to have the best potential for meeting the conditions necessary for realizing the ultimate ultrahigh sensitivity imaging device, i.e. the conditions for obtaining a high S/N ratio at the theoretical limit.

To achieve the ultimate ultrahigh sensitivity imaging device, the following three conditions have to be met:

(i) All of the incident photons must be guided to the photoelectric conversion part (100% fill factor).

(ii) All the photons must be converted into electrons in the photoelectric conversion part (100% photoelectric conversion efficiency).

(iii) It must be possible to amplify the converted electronic signal without adding any noise.

A camera tube has a fill factor of 100% and thus satisfies condition (i). Also, because the target uses the internal photoelectric effect, it is also easier to increase the photoelectric conversion efficiency than in imaging devices based on the external photoelectric effect, such as image devices using intensifiers. In other words, it can also satisfy condition (ii). Accordingly, if a way can be found to satisfy condition (iii), then it will be possible to obtain imaging devices with unparalleled sensitivity and picture quality. To achieve this, it is first necessary to bring about some form of amplification within the target. Based on this reasoning, NHK began to research targets with in-built amplification capabilities, as described later.

Target types and amplification effects

As described earlier, targets can be classified into two types: an injection type where electrical charge is injected into the film from outside, and a blocking type where the injection of electrical charge is blocked. In the injection type, an amplification effect is obtained whereby the external circuitry extracts a greater number of electrons than the number of photons incident on the target. Although blocking types result in good picture quality with low lag and low dark current, it has not been possible to produce amplification effects in such targets. We therefore concentrated our studies on injection type targets.

However—to cut a long story short—the HARP camera tube target is actually a blocking type, not an injection type. Injection type targets suffer from drawbacks such as a susceptibility to increased dark current and the amplification of lag by the same gain. However, the reason why we applied ourselves to the study of injection-type targets is because at that time there were thought to be no other ways of conferring amplification properties to the target.

Although NHK's research initially focused on injection-type targets for these reasons, in 1985 an unusual experiment was conducted involving making a blocking-type target behave like an injection-type target by forcibly applying a large voltage. This led to the discovery of a phenomenon whereby the sensitivity is increased within the photoelectric conversion film in a manner that could not be explained

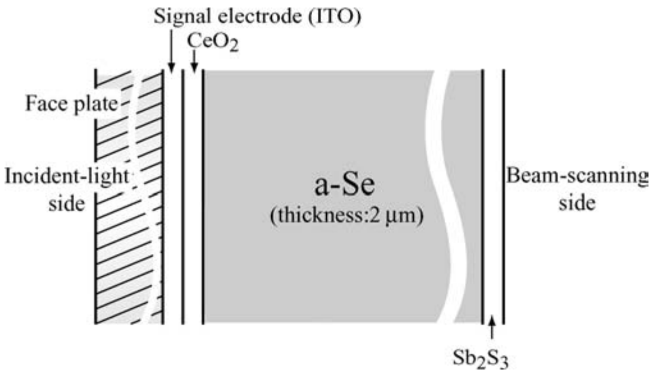


Figure C2.1.7. Structure of prototype target.

in terms of charge injection. This was the starting point for the development of the HARP camera tube. The experiment is described below.

Experimental details

At the time, NHK was working on an injection-type target using a-Se, which can produce an amplification effect with a relatively weak electric field ($5 \times 10^6\,\mathrm{V\,m^{-1}}$). A weak electric field reduces the efficiency with which photons are converted into electrons, and thus gives rise to problems such as increased shot noise.

It was deduced that this noise could be reduced by subjecting the target to a suitably strong electric field before injecting the charge, so the target was initially designed for an experiment to inject electrons as shown in figure C2.1.5 by forcibly applying a very high voltage to a blocking type target that has a structure in which the injection of charge is blocked. Figure C2.1.7 shows the structure of the prototype target we produced for this experiment. The photoconductive film in this target is a 2-μm-thick a-Se film formed by vacuum deposition (vacuum: 1.33×10^{-4} Pa). The target is of the blocking type. Like the Saticon, it blocks the injection of holes at the junction formed between the a-Se film and the transparent signal electrode (ITO: indium–tin oxide) and CeO_2 layers, and blocks the injection of electrons through the use of an Sb_2S_3 layer. But, in contrast to the Saticon, it does not include high-concentration Te- and As-doped layers to concentrate an electric field near the signal-electrode interface, which means that even better hole-injection-blocking characteristics can be expected. Also, for the Sb_2S_3 layer, inert-gas (Ar) pressure at the time of deposition was set to 31.9 Pa considering the porous-film fabrication conditions that would suppress the emission of secondary electrons even when target voltage is exceptionally high and promote stable low-velocity beam landing. The thicknesses of the CeO_2 and Sb_2S_3 films in the target are 20 and 100 nm, respectively, indicating that these two films are considerably thinner than the Se layer. Target thickness can therefore be regarded as essentially the same as that of the Se layer.

Current–voltage characteristics

Figure C2.1.8 shows target current–voltage characteristics of the HARP tube I. Blue light (centre wavelength: 440 nm) is used here as incident light. From figure C2.1.8, we see that signal current increases rapidly as target voltage increases from 0 V but comes to saturation, at least temporarily, starting at about 20 V. This saturation region is thought to correspond to the state where most electron–hole pairs

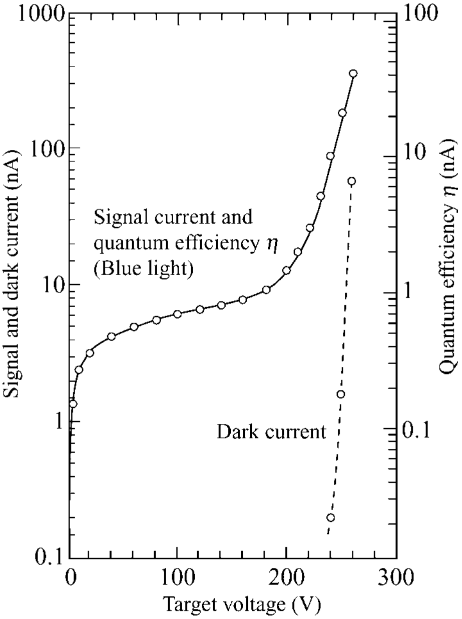


Figure C2.1.8. Signal current and dark current *versus* target voltage.

excited in the a-Se film by incident light have come to separate under a strong electric field within the film becoming signal current as a result. As target voltage continues to increase, however, we see the phenomenon of signal current again rising dramatically beyond this saturated region.

Quantum efficiency η with respect to blue light in a-Se film is estimated to be 0.9 for an operating electric field of $8 \times 10^7\,\mathrm{V\,m^{-1}}$ [8]. In the HARP tube, this electric-field strength corresponds to a target voltage of 160 V, and this fact enables us to establish a scale for η on the right vertical axis in figure C2.1.8.

This scale tells us that η exceeds 1 at a target voltage of 180 V and reaches 10 at 240 V. Furthermore, at a target voltage of 260 V, η is 40 and extremely high sensitivity occurs in the HARP tube I. As for dark current, it also becomes large in the high-voltage region, but it is nevertheless quite small at 0.2 nA under target-voltage operating conditions of 240 V ($\eta = 10$). As described earlier, the phenomenon of increased sensitivity with η exceeding 1 has been observed when operating an a-Se photoconductive target with a blocking-type structure in a very strong electric field.

The origin and operating principle of the HARP camera tube

Since η was greater than 1 in the prototype target, it was thought that an amplifying action occurred due to the expected blocking type target behaving as an injection type. However, we found that this target exhibited hardly any dependence of lag on the applied voltage, which one would normally expect to see in an injection type target. Specifically, according to the operating theory of injection type targets, the effective storage capacitance of the film increases by an amount corresponding to the magnitude of the gain, so when the applied voltage exceeds 180 V the lag ought to increase steeply. However, such phenomenon was seen in the test target. We therefore performed several new experiments. As a result, it became clear that the phenomena exhibited when $\eta > 1$ conform to the following properties:

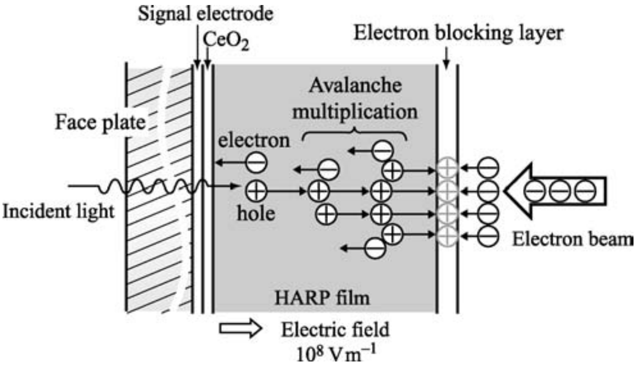


Figure C2.1.9. Operating principle of the HARP target.

(i) The effective storage capacitance of the target is constant and does not increase even when η is greater than 1.

(ii) The amplification effect is dependent on the direction of the incident light, and compared with the face plate side, the degree of amplification is smaller when light is incident from the beam scanning side. That is, the target has a higher gain for hole transport than for electron transport.

(iii) When the electric field inside the target has a constant intensity, the amplification gain increases as the a-Se layer gets thicker.

Based on these findings, it was concluded that the amplification action obtained with the test targets is not due to the injection of charge but is due to a avalanche amplification effect that occurs stably and continuously in blocking type targets for imaging devices. This marks the origin of HARP camera tubes that use the avalanche amplification phenomenon, and in this way HARP camera tubes were born out of research into completely different targets where charge injection activity is taken into consideration.

Figure C2.1.9 schematically illustrates the operating principle of this target. Electrons and holes produced by the incident light are accelerated inside the target, to which a strong electric field of about $10^8\,\mathrm{V\,m^{-1}}$ is applied, and new electron–hole pairs are then generated successively by impact ionization. As a result, a large number of electrons are extracted from the signal electrode for each incident photon. The high sensitivity of HARP camera tubes is due to the avalanche multiplication effect in the a-Se target and the fact that this multiplication results in hardly any added noise. Furthermore, a HARP camera tube also has superior lag characteristics and resolution as mentioned later.

Basic structure of targets for practical use

Figure C2.1.10 shows the basic structure of a HARP camera tube target for practical use. Like the prototype target as shown in figure C2.1.7, it uses layers of a-Se, CeO_2 and Sb_2S_3. However, the target for practical use also contains arsenic (As), lithium fluoride (LiF) and tellurium (Te). The arsenic suppresses crystallization of the a-Se, thereby preventing the generation of defects. The lithium fluoride serves to control the electric field inside the target, and prevents the generation of defects by decreasing the electric field near the interface between the a-Se film and the CeO_2. The tellurium increases the target's sensitivity to red light, and is added to the target for the red channel. The parts to which LiF and Te are added are exceedingly thin, and are no more than a few per cent of the overall target film thickness.

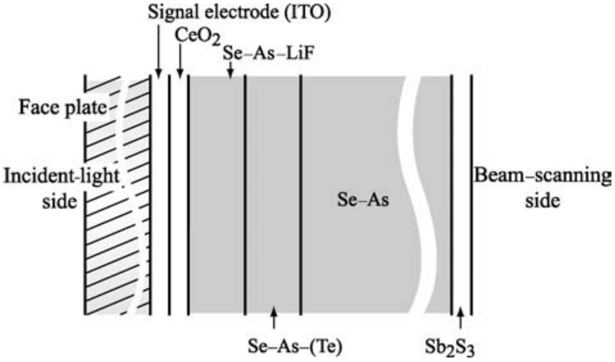


Figure C2.1.10. Structure of the HARP target.

C2.1.1.5 The evolution of HARP camera tubes

When HARP camera tubes were first developed, they had a target film thickness of 2 μm, and their sensitivity was about 10 times that of conventional Saticon camera tubes (figure C2.1.11). By taking advantage of the fact that these targets also have high resolution, HDTV camera tubes were developed which were put to use at the Seoul Olympics [9]. But as they came to be used increasingly for TV programmes, a demand arose for even higher sensitivity. Furthermore, since CCDs—which had by then become the most common imaging devices—have virtually no lag problems, there was also a demand for improving the lag characteristics of HARP camera tubes.

As can be seen from the operating principle shown in figure C2.1.9, the avalanche multiplication factor of a HARP camera tube increases as the target gets thicker, resulting in greater sensitivity. The lag also decreases as the film thickness increases. This is due to reduction of the target storage capacitance, which dominates the lag characteristics. Consequently, by increasing the thickness of the layer consisting primarily of a-Se in the target to around 6–8 μm, it was able to develop a practical HARP camera tube with improved lag characteristics that had 60–80 times the sensitivity of a Saticon tube.

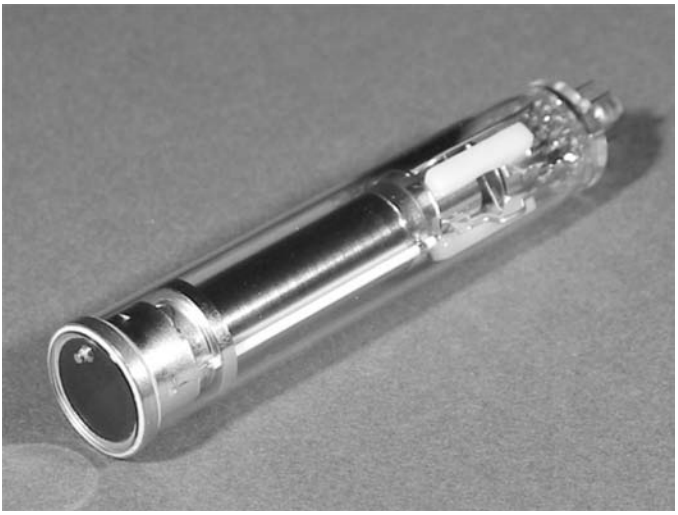


Figure C2.1.11. Appearance of the ultrahigh sensitivity HARP camera tube.

Table C2.1.1. The evolution of HARP camera tubes.

	Year in which developed		
	1985	1990	1995
Target film thickness (μm)	2	6–8	25
Sensitivity (relative to a Saticon tube)	About 10×	60–80×	About 600×
Lag (50 ms)	4.6%	1.5–1.2%	Below measurable limit (theoretical value: 0.09%)

Furthermore, experiences such as news coverage of the Kobe earthquake disaster in Japan (1995) resulted in increased demand for the development of an ultrahigh sensitivity imaging device capable of producing, for example, aerial night-time shots of the stricken region which had been plunged into darkness due to power failures. NHK, Hitachi, Ltd, and Hamamatsu Photonics K.K. therefore studied ways of making the HARP camera tube even more sensitive, as described later, and developed an ultrahigh sensitivity HARP camera tube with the target film thickness increased to 25 μm whose sensitivity is 600 times greater than that of a Saticon tube. The lag of this camera tube was reduced to a level below the measurable limit. Note that since modern CCDs are about six times as sensitive as Saticon tubes, the sensitivity of this HARP camera tube is about 100 times greater than that of a CCD.

Table C2.1.1 shows how the target film thickness, sensitivity (relative to a Saticon tube) and lag (the value 50 ms after the incident light is cut off) of HARP camera tubes have changed over the years. In the following, characteristics of the HARP camera tube with a target film thickness of 25 μm, whose sensitivity exceeds that of the naked eye, will be mentioned.

C2.1.1.6 Principal characteristics of the ultrahigh sensitivity HARP camera tube

This section describes the principal characteristics of the ultrahigh sensitivity HARP camera tube with a target film thickness of 25 μm (2/3 in MM type, shown in figure C2.1.14) [10, 11].

Sensitivity

Figure C2.1.12 shows how the signal current (which represents the sensitivity) and dark current vary with the applied voltage. By way of comparison, this figure also shows the signal current measured using an ordinary (Saticon) camera tube subjected to the same amount of incident light. With an applied voltage of 2500 V, the HARP camera tube is over 600 times more sensitive than the Saticon tube. The dark current in this case is about 2 nA.

Note that since the HARP camera tube's sensitivity can be varied greatly by controlling the applied voltage, it can be adjusted to the sensitivity of ordinary imaging devices by reducing this voltage. In other words, it can also be used to take pictures in very bright situations such as daylit outdoor scenes.

Spectral response characteristics

Figure C2.1.13 shows the spectral response characteristics of an ordinary HARP target and a Te-doped target with increased sensitivity to red light (for use in the red channel). In the a-Se layer of the HARP target, there is little cancellation of charge due to recombination, even close to the junction interface with

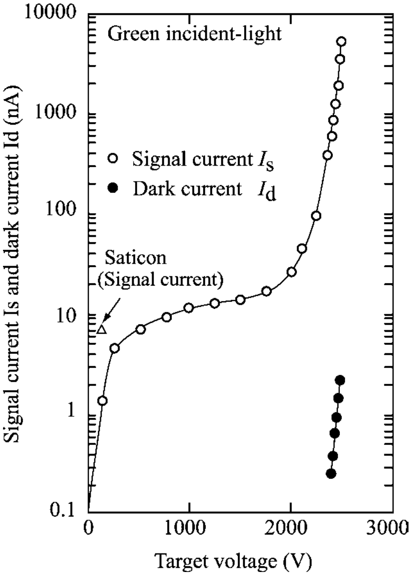


Figure C2.1.12. Signal current and dark current *versus* target voltage in the ultrahigh sensitivity HARP camera tube.

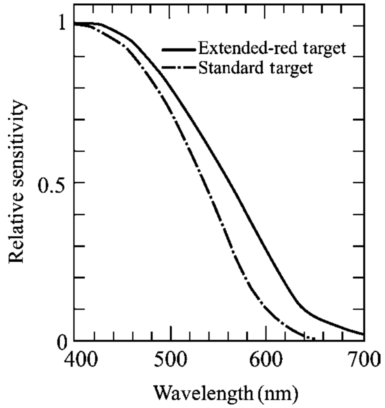


Figure C2.1.13. Spectral response characteristics of the ultrahigh sensitivity HARP camera tube.

the CeO_2, so a high photoelectric conversion efficiency is obtained for short-wavelength (blue) light that is absorbed in this part. On the other hand, since the band gap of a-Se is about 2.0 eV, the limit of sensitivity to light of longer wavelengths (red light) is the corresponding wavelength which is about 620 nm (standard type). Since the red channel of a colour camera should exhibit sensitivity up to about 700 nm [12], the target for the red channel is made more sensitive to red light by doping it with Te.

Lag characteristics

The lag characteristics are determined by the storage capacitance of the film and the electron-beam temperature of the scanning electron beam. When the target film thickness is 25 μm, the theoretical value of the lag in the third field after turning off the incident light is 0.09%. This value is calculated from a target-layer storage capacitance of 130 pF and an electron-beam temperature of 3000 K. As shown in the right-hand column of table C2.1.1, the lag of the HARP camera tube is below the measurable limit when the target film thickness is 25 μm.

Resolution characteristics

The HARP camera tube has a limiting resolution of more than 800 TV lines, and no degradation of resolution due to the avalanche multiplication action was observed. Since the resolution characteristics of the camera tube are controlled by the spot diameter of the scanning electron beam, even higher resolution can be obtained by combining with electron optics having a smaller beam spot diameter.

Noise

The magnitude of the noise added as a result of avalanche multiplication is represented by the excess noise factor, but in the case of a HARP camera tube its value is approximately 1. In other words, the amplification achieved with this camera tube is almost noise-free. The reason why this result is obtained is thought to be because of the large ratio of the respective ionization coefficients of holes and electrons in the a-Se film and because the beam scanning side of the photoelectric conversion film is in a floating state (its potential is not fixed), whereby a type of negative feedback action takes place which controls the noise.

In this way, the HARP camera tube combines the characteristics needed for high sensitivity with the characteristics needed for superior picture quality.

C2.1.1.7 Ultrahigh sensitivity HARP cameras and their applications

A hand-held HARP colour camera has been produced using HARP tubes with a 25 μm thick target. Its appearance is shown in figure C2.1.14. Table C2.1.2 shows major specifications of the camera. This three-tube colour camera with the target voltage of 2500 V has greater sensitivity than the naked human eye and can obtain clear colour images even in lighting conditions equivalent to moonlight. Figure C2.1.15 shows an example of how the images taken with such a camera compare with the images

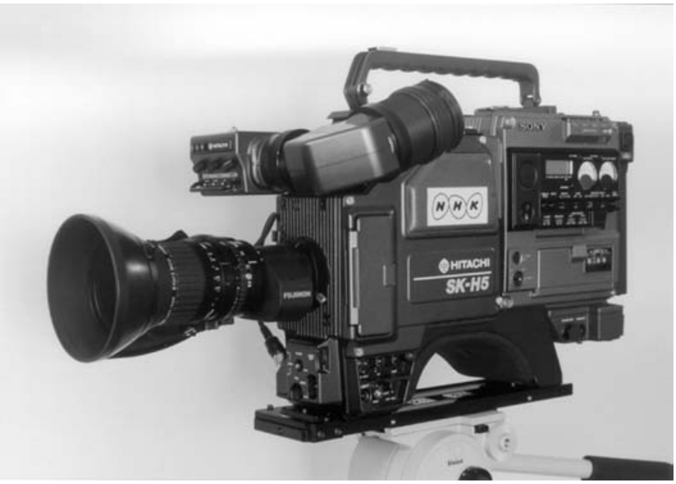
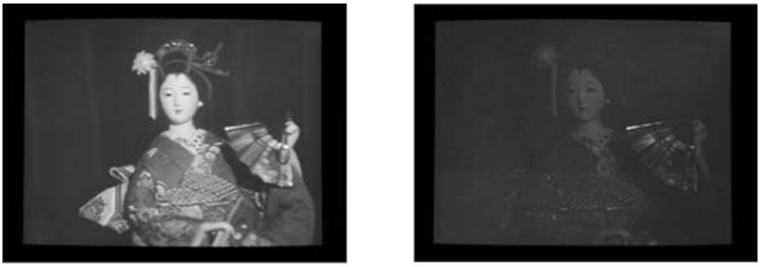


Figure C2.1.14. Appearance of the ultrahigh sensitivity HARP colour camera.

Table C2.1.2. Specifications of the ultrahigh-sensitive HARP camera.

Maximum sensitivity	11 lux, F8
Minimum scence illumination	0.03 lux (F1.7, +24 dB)
Signal-to-noise ratio	60 dB
Resolution	700 TV lines
Amplifier gain selection	0, +9 dB, +24 dB
Weight	5 kg
Power consumption	25 W



(*a*) Image taken with the HARP camera

(*b*) Image taken with a CCD camera (+18 dB)

Figure C2.1.15. Monitor pictures produced by colour cameras with the ultrahigh sensitivity HARP tubes and CCDs. Illumination is 0.31× and lens irises are at F1.7.

taken by a CCD hand-held camera under the same conditions (subject illumination: 0.3 lux, lens aperture: F1.7). The dark subject was difficult to view with the CCD camera even when its gain was boosted by 18 dB, whereas the HARP camera was able to produce a clear image. It was confirmed by the colour camera test that the HARP camera is about 100 times as sensitive as the CCD camera. It goes without saying that the sensitivity of the HARP camera can be decreased by decreasing the target voltage, so that the camera is capable of producing excellent picture quality over a wide range of shooting conditions from daylight to moonlight. In addition, the HARP tube offers other excellent features, such as insensitivity to burning, compared with the Saticon camera tube. It was also confirmed that the additional noise produced by the avalanche multiplication was negligibly small.

The HARP camera is a powerful tool in reporting breaking news at night and the production of scientific programmes. In addition, since a HARP target can convert x-rays into electrons directly, it should be possible to exploit this capability to produce x-ray imaging devices with unparalleled levels of resolution and sensitivity. Consequently, this technology is attracting high levels of interest for applications such as the early detection of cancer and diagnosis of heart disease.

C2.1.1.8 Conclusion

This section has provided a description of the mechanism of the photoconductive camera tube covering past types and the ultrahigh sensitivity HARP camera tube based on a new operating principle.

In general, ultrahigh sensitivity imaging devices have so far been regarded as special-purpose devices, in a separate class to ordinary high picture quality imaging devices such as CCDs. But if we can develop the ultimate ultrahigh sensitivity imaging device—with noise-free internal amplification and extremely high gain, and whose fill factor and photoelectric conversion efficiency are 100%—then we will have a device with ultrahigh picture quality and an S/N ratio close to the theoretical limit. Such a device will be able to take clear pictures with less noise than any other existing device under all lighting conditions. Although HARP camera tubes themselves are coming very close to this ultimate goal, we would still like to make further improvements to the photoelectric conversion efficiency.

C2.1.2 CCD and CMOS imaging sensors

C2.1.2.1 Introduction

Research and development work involving solid-state image sensors was started in 1960s [13, 14]. At the beginning, an MOS image sensor went ahead and the first video camera using it was put to practical use in 1981, because the MOS devices could be fabricated by relatively simple technology similar to that of the DRAMs. A charge-coupled device (CCD) was invented by Boyle and Smith [15]. Since the preliminary experimental result that a charge transfer phenomena along a surface of the CCD could be applicable to a scanning operation essential to a solid-state image sensor was recognized, application of the CCD to the solid-state image sensor was accelerated all over the world. The CCD image sensor using a charge-coupled device concept is now widely employed in video cameras, digital still cameras and so on. The main reason is that the CCD image sensor has a high signal-to-noise performance, although a somewhat complex technology is required in fabricating the CCD.

Advantages of the CCD image sensor, compared with the image pick-up tubes, which are exclusively used for TV cameras up to now, are:

- higher reliability;
- uniform spatial resolution;
- lower supply voltage and power consumption;
- smaller in volume and weight;
- lower in cost;
- no image lag.

In 1994, an active imaging sensor fabricated with a CMOS-LSI compatible technology, which was especially driven by the technology push of the DRAMs, was proposed by Mendis *et al* [16]. Advantages of the CMOS image sensors are much lower supply voltage, less power, much lower cost and faster readout rate than that of the CCD image sensor. Moreover, the sensor is ideal for implementing on-chip signal processing circuits together with an image sensing circuit. Consequently, the CMOS image sensor will be useful in machine vision, smart sensor and various mobile camera applications.

To efficiently convert two-dimensional light information to a successive video signal corresponding to it, the image sensor has to provide the three functions as follows:

- photon sensing in a pixel;
- integration of photo-induced charge;
- readout of the integrated charge and device architecture convenient to fulfil a parallel to serial conversion.

C2.1.2.2 Photon sensing

Solid-state imaging is based on the principle of converting incoming light to electron charges at the input of an image sensor and the electron charges to signal voltages at the output of the device. To do this the photon conversion, and the transport and detection of the signal charges, are greatly important.

Photon conversion

Photons falling on a pixel array of the solid-state image sensor are absorbed if their energy is larger than that of the band-gap of the silicon substrate and causes the generation of electron–hole pairs in silicon. The generated electrons and holes are separated by the electric field applied on the pixels, and then the electrons are collected into the pixel and holes are drained into the substrate in the case of the p-type substrate.

As the photons are generally incident on the sensitive area in the pixel of a solid-state image sensor through a thin film, for example SiO_2 or Si_3N_4 and others, or semitransparent electrode of poly-silicon film, the fluctuation in the spectral sensitivity distribution of the sensor appears. Figure C2.1.16 shows a calculated result of the spectral transmission through a $SiO_2(d_1)$–poly-Si(d_2)–$SiO_2(d_3)$ structure deposited on the silicon wafer. This means that the wavelength, at which the maximum or minimum can be observed in the spectral transmission, is dependent on the thickness and refractive index of these films, respectively. The refractive indexes of the materials commonly employed for the solid-state image sensors are shown in table C2.1.3.

Charge integration

If the electron charges collected in a pixel for a short period only while the pixel is selected, are read out and transported to an output stage, the signal level at the output stage is much too small to convert it into a detectable voltage. This problem could be avoided by locally integrating the light-induced electrons in each pixel. This action was proposed by Weckler [17], and called a charge integration mode. In this operation, the generated electrons are separately stored in each pixel during a certain amount of time and when the pixel is selected again, the stored charge packet is taken out. Consequently, the number of electrons in a charge packet is increased in proportion to the integration times and can be large enough to convert it into the signal voltage which is easily detectable at the output stage.

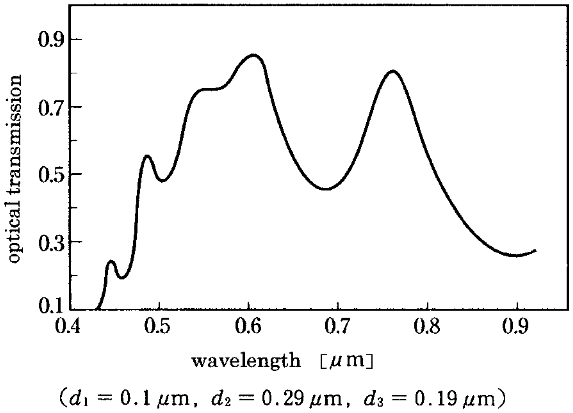


Figure C2.1.16. Spectral transmission curve calculated under SiO_2 (d_1)–poly-silicon (d_2)–SiO_2 (d_3) system.

Table C2.1.3. Refractive indexes of the key materials.

Materials	Refractive index
Air	1
Poly-glass	1.46
Poly-Si, Si	$3.42 - j\left(\frac{\alpha\lambda}{2\pi}\right)$
SiO_2	1.46
Si_3N_4	~2
In_2O_3	1.97

α is absorption coefficient and λ is light wavelength.

In aid of the basic idea, the solid-state image sensor serviceable to the various video camera applications could be realized.

The charge integration is easily actualized by introducing a small capacitor in each pixel. Two typical organizations are shown in figure C2.1.17. One is a metallurgical np-junction reverse biased and the other is a voltage-induced np-junction so-called MOS capacitor. In the metallurgical junction, when an MOS switch is on by applying the positive pulse voltage to the gate electrode, the np-junction is reverse biased. After the pulse is off, the np-junction is electrically isolated and thus remains reverse biased until the following pulse is applied to the gate again. All of the electrons photo-generated during this interval are collected and then integrated in the np-junction capacitance. The voltage across the capacitor decreases accordingly with increase of the numbers of electron stored in the capacitance.

In the MOS capacitor, a positive bias voltage is always applied on a gate of the MOS capacitor. Consequently, when the positive pulse is applied to the MOS switch, the MOS capacitor is reverse biased and the integration of the electrons takes place in the similar manner of the np-junction.

Charge transport and detection

Whenever the selection pulse is applied, the stored charge packet is readout from each pixel and then carried towards the output terminal of the sensor. To achieve this, two alternatives as shown in figure C2.1.18 are well known. One is an MOS switch array with a sense line and the other is a CCD shift register. In the former the MOS switch is in turn on and the stored charge packet is accordingly read out from the selected pixel. An advantage of this method is that time lag between both the timings of pixel selection and corresponding video signal detection at output of the sensor can be minimized. But the MOS switch connects small capacitance of the pixel to the relatively large capacitance of the sense line. Thus, the signal-to-noise performance of this construction is rather poor.

In the latter all the pixels are selected at the same time and the charge packets in the pixels are moved to the shift register in parallel. After that, these packets are transferred towards the output diffusion by appropriate clocking. The detection of the charge packets are done by converting it to a voltage on the floating diffusion followed by a source follower [18] as shown in figure C2.1.19. This charge detection circuit calls a floating diffusion amplifier (FDA). The output voltage ΔV of the FDA is given by the following equation:

$$\Delta V = A_v \Delta V_A = \frac{Q_s}{C_A} \frac{g_m}{1 + g_m R_s} \tag{C2.1.1}$$

where Q_s is the amount of charge flowing to the floating diffusion, C_A the capacitance of the floating diffusion, A_v the voltage gain of the FDA and g_m the transconductance of the FDA. A typical value of the sensor conversion gain $\Delta V / Q_s$ is now about 10–20 $\mu V\,e^{-1}$.

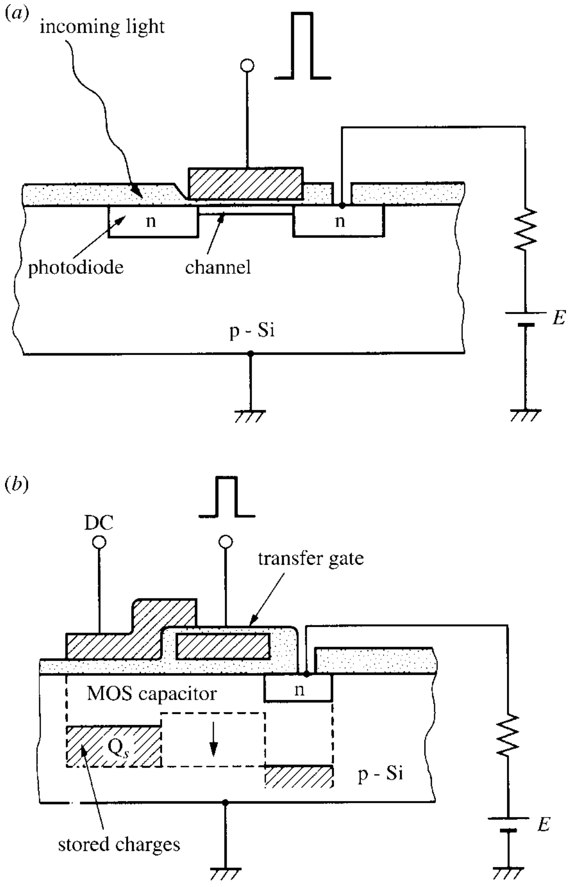


Figure C2.1.17. Photo-sensing and charge integration: (*a*) a metallurgical np junction and (*b*) a voltage induced junction.

The charge packet is moved from the small pixel capacitance to the small output diffusion capacitance. This construction with a CCD shift register, therefore, results in the relatively high signal-to-noise performance, but the difference between both the timings of readout of the integrated charge packet from the pixel and detection of the same packet at the output of the device, is relatively large and differs from pixel to pixel.

C2.1.2.3 Device architecture of CCD image sensor

Linear image sensor

A linear image sensor has only a scanning scheme in a direction along a row of light-sensitive elements. Therefore, in order to read-in the two-dimensional objects such as various documents, it is necessary to move the object in the direction perpendicular to the scanning one. A device having a large number of elements, applicable to copy and facsimile machines, etc is required.

The simplest linear image sensor is composed of a single line of light-sensitive elements, such as photodiodes or MOS capacitors, and a CCD shift register located next to the row of the pixels, needed for the readout of the charge packets photo-generated in each pixel. The pixels are isolated from the

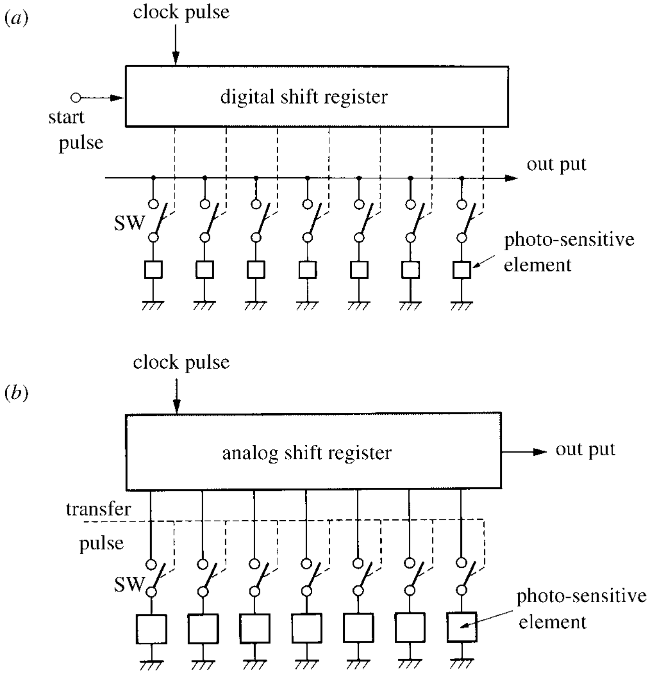

(a)
clock pulse
digital shift register
start pulse
out put
SW
photo-sensitive element
(b)
clock pulse
analog shift register
out put
transfer pulse
SW
photo-sensitive element

Figure C2.1.18. Readout structure of signal charge: (*a*) an MOS switch and (*b*) a CCD shift register.

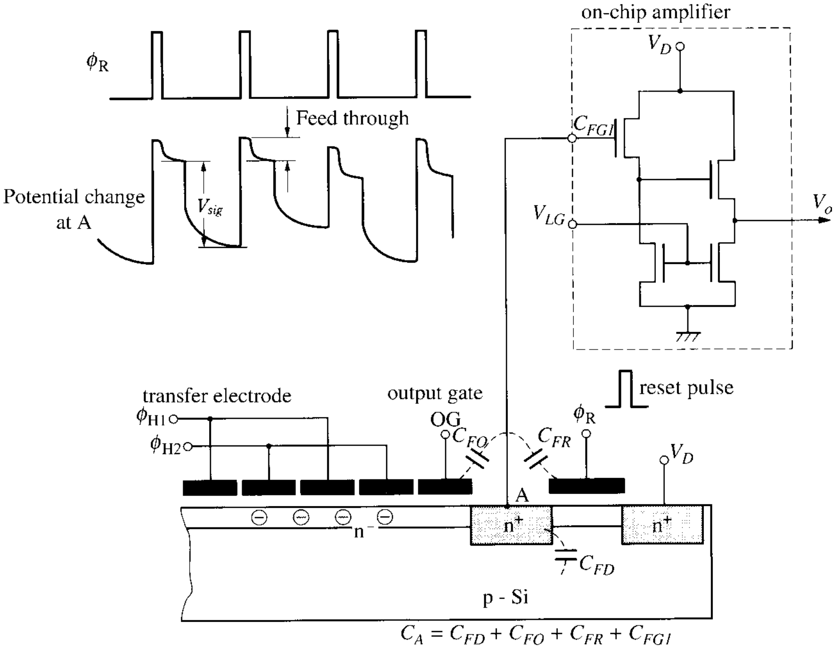

on-chip amplifier
ϕ_R
V_D
Feed through
C_{FGI}
Potential change at A
V_{sig}
V_{LG}
V_o
reset pulse
transfer electrode
output gate
ϕ_{H1}
ϕ_{H2}
OG
C_{FO}
C_{FR}
ϕ_R
V_D
A
n^+
n^+
n^-
C_{FD}
p - Si
$C_A = C_{FD} + C_{FO} + C_{FR} + C_{FGI}$

Figure C2.1.19. Floating diffusion amplifier and related signal output.

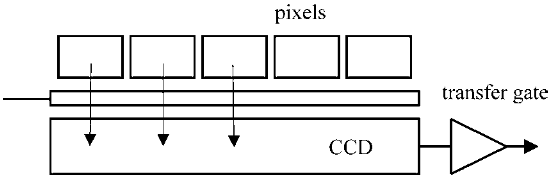


Figure C2.1.20. Device concept of a simple linear image sensor.

CCD shift register by a transfer gate, as shown in figure C2.1.20. After the integration of the charge carriers in each pixel, a high pulse voltage, enough to switch on the gate, is applied to the transfer gate and then the charge carriers are moved into the CCD shift register one at a time. Switching off the transfer gate, a new integration of the photo-generated charges is started. During the integration, the charge packets located in the CCD shift register are transferred through the CCD shift register towards the output of the device in turn and are converted to the voltages proportional to the charge packet size. By this configuration the conversion of spatial information to a time-varying signal can be realized.

The highest resolution of the linear image sensor is restricted by the smallest interval of the CCD unit cell, because the interval of the pixel is equal to that of the CCD unit cell. Note that the CCD shift register has to be shielded from the incoming light to avoid a mixing of the transfer charges with the photo-induced charges in the shift register.

In a bilinear CCD image sensor, the number of the pixels per unit length can be increased to twice as many as that of the normal linear sensor, because it is possible to halve the interval of the pixel compared with that of the CCD unit cell as shown in figure C2.1.21. Therefore, a high-resolution image sensor is easily realized by using the bilinear structure with two CCD shift registers, which are located on each side of the row of light-sensitive elements.

After the photo-generated charges are integrated for a given period of time, the charge packets caught in each pixel are moved to two CCD shift registers by application of the pulse voltage to the transfer gates. For example, odd pixels are moved to the bottom register and even pixels to the top register. At the output of the two registers the charge packets are converted to electrical signals, respectively. Then the electrical signals are multiplexed by reading these one after the other and joining them in the shape of a successive video signal. To avoid a spurious signal, a so-called fixed pattern which appears in a reproduced picture, it is necessary to exactly compensate for a variance of the performance of the two output amplifiers.

In a copying machine or facsimile application, two kinds of an optical system are usually employed. One is a focusing lens system, by which the image on a document is reduced and projected on a surface of the sensor as shown in figure C2.1.22(*a*). An advantage of this system is low sensor cost but it may be

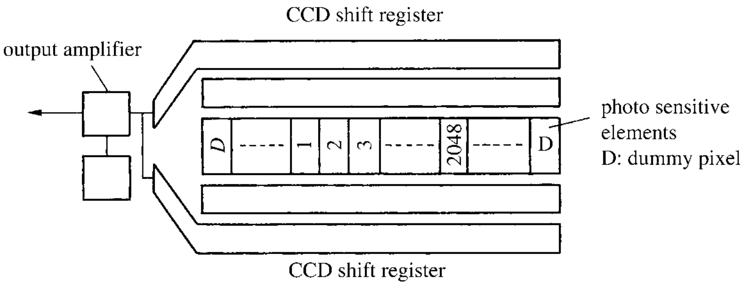


Figure C2.1.21. Construction of a bilinear image sensor.

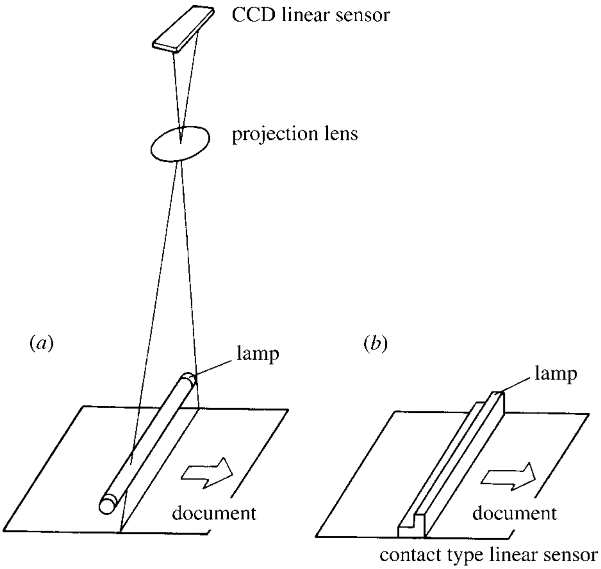


Figure C2.1.22. Optical systems: (*a*) focusing lens type and (*b*) direct contact type.

difficult to miniaturize the machine, because of a finite optical path length of the lens system. An other is a thin fibre optics plate system, for instance a selfoc lens array, mounted on the sensor surface. The image on a document is projected by directly contacting the document as shown in figure C2.1.22(*b*). To realize it, the length of a row of the light-sensitive elements must be longer than the width of the document. The length of a single linear sensor, fabricated using a crystalline silicon wafer, has usually been restricted to a few centimetres, to minimize the cost. Such a contact type linear sensor is formed by connecting several sensor chips in series. In designing this multi-chip linear sensor, it is important to minimize the degradation of a resolving power at the joint where the chips are connected together. At present, the multi-chip linear sensor, which joins the several chips in line or zigzag as shown in figure C2.1.23, has been developed.

The typical performance of the multi-chip linear sensor joining 14 chips together with an accuracy of 6 μm is presented in table C2.1.4.

The linear sensor fabricated by poly-silicon thin film technology is available too. The device has advantages: low cost and no restrictions in the length of the sensor.

Area image sensors

The typical area image sensor to convert a two-dimensional optical image to a successive video signal can be roughly grouped into five device architectures: interline transfer CCD, full frame transfer CCD, frame transfer CCD, frame–interline transfer CCD and X–Y addressed CMOS.

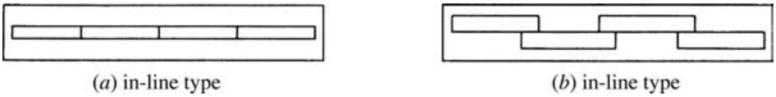


Figure C2.1.23. Multi-chip linear image sensors: (*a*) in-line device and (*b*) zigzag device.

Table C2.1.4. Typical performances of the multi-chip linear imaging sensor.

Chip size	Effective length	Resolution	Readout speed	Operation voltage	Power consumption
$16.0 \times 0.74\,\mathrm{mm}^2$	244 mm (A4)	8 dots/mm	5 ms/line	5 V	50 mA (sensor) 320 mA (lamp)

Interline transfer CCD

An interline CCD image sensor is formed by arranging several linear CCD imagers in parallel. The typical construction of the device is shown in figure C2.1.24. The light-sensitive elements, which are photodiodes or MOS capacitors, are located near the vertical CCD shift registers. The horizontal output register is connected to the final stages of the vertical shift registers. The surface of the vertical shift register and horizontal shift register is covered by opaque layers (Al, W–Si, etc) through a proper thin insulator film to shut off the incoming light.

The basic operation of the devices is as follows: the charge packets stored in each pixel are simultaneously moved from the pixels to the neighbouring vertical shift registers by applying an interrogating pulse on the transfer gate. After this transfer, the pixels begin to integrate the photo-generated charges again. The charge packets located in the vertical shift registers are transferred downwards in parallel by proper clocking and then they are carried into the horizontal shift register line by line. The charge packets in the horizontal shift register are quickly transferred to an FDA and are

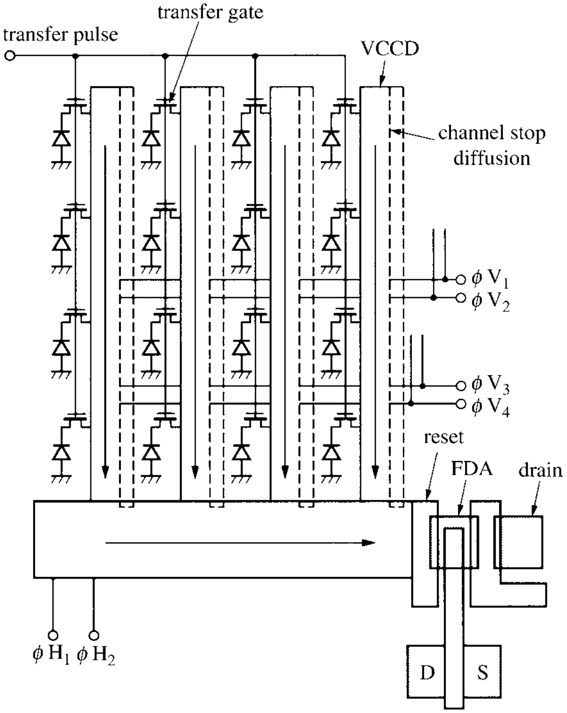


Figure C2.1.24. Device construction of an interline transfer CCD image sensor.

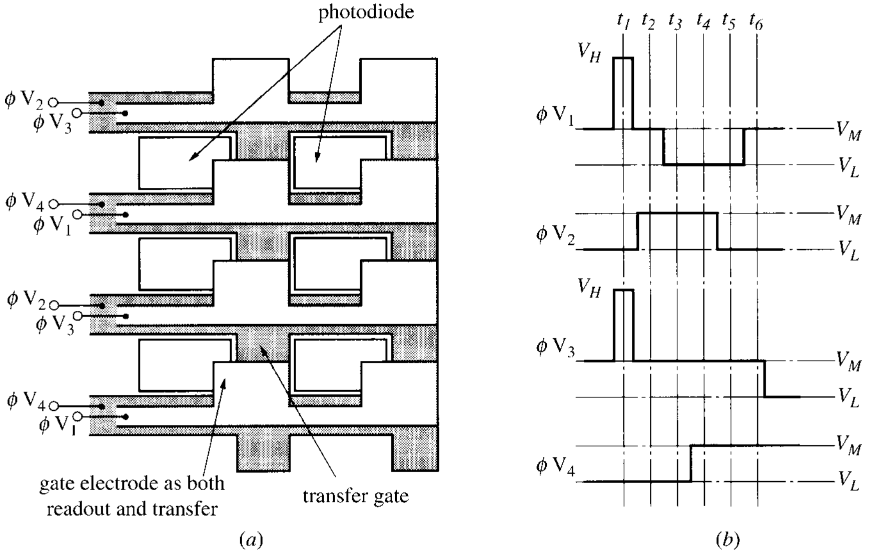

photodiode
ϕV2
ϕV3
ϕV4
ϕV1
ϕV2
ϕV3
ϕV4
ϕV1
gate electrode as both readout and transfer
transfer gate
(a)
t1 t2 t3 t4 t5 t6
VH
ϕV1
VM
VL
ϕV2
VM
VL
VH
ϕV3
VM
VL
ϕV4
VM
VL
(b)

Figure C2.1.25. (*a*) Pixel layout in an interline transfer CCD. (*b*) Four phase clocking scheme.

converted into an electrical voltage. This parallel to serial transfer is similar to that described for the linear image sensor.

The typical pixel configuration used in many interline transfer CCD image sensors for consumer applications is shown in figure C2.1.25 together with the four phase clock pulse sequence applied to the transfer gates of the vertical shift register. In this device one of the four gates, which is formed by self-aligning the gate to an edge of the pixel, can be used both as a charge transfer electrode during a charge transfer period and as a readout-switch of the charge packet during a readout period.

Schematic potential profiles in depth in the CCD, corresponding to the clock pulse voltages of V_H, V_M and V_L are shown in figure C2.1.26. When the pulse voltage V_H (= +15 V) is applied to the transfer gates Φ_1 and Φ_3 at the time t_1, a potential barrier between the photodiode and the vertical shift register

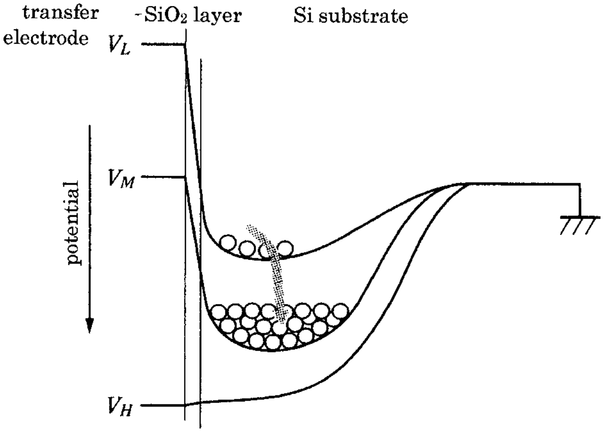

transfer electrode
-SiO2 layer
Si substrate
VL
VM
VH
potential

Figure C2.1.26. Potential profiles under the transfer gate, where three clock pulse voltages of V_H, V_M and V_L are applied.

disappears and the charge packets stored in the photodiodes flow into the vertical shift register through the channels under the gates Φ_1 and Φ_3, and then the photodiodes begin to store the photo-generated charges again. At the time t_2, two adjoining charge packets are summed to achieve the field integration corresponding to the interlace scanning for TV applications. At the time t_3, the pulse voltage applied to the gate Φ_1 is decreased to V_L ($=-5\,\mathrm{V}$), the charge stored under the gate Φ_1 flows out into a depletion region under the adjacent gate Φ_2, where the pulse voltage of V_M ($=0\,\mathrm{V}$) is applied. Similar charge transfer is also repeated under the transfer gates Φ_3 and Φ_4, and the charge packet sequence is transferred down the horizontal CCD one after another. In the horizontal CCD the charge packets are serially transferred to the FDA and are converted into an electrical voltage (video signal) in a similar way to that described in section C2.1.3.

For a high-definition TV application, a transfer speed of the horizontal shift register higher than about 74 MHz is required. In this case, the register driven by two-phase pulse clocking has been widely employed. The structure of the horizontal shift register for two phases pulse clocking and the sequence of the charge transfer are illustrated in figure C2.1.27. In the two-phase pulse clocking, the scheme that intercepts a buck flow of the charge packet is indispensable. To do this, potential barriers are periodically laid along the charge transfer channel as shown in the figure. These barriers are formed by ion-implantation of p-type dopants, such as boron, etc through the first poly-silicon layer previously deposited and patterned. After that, second poly-silicon electrode is formed and two adjacent electrodes are mutually connected as shown in the figure.

Recently, it has become possible to fabricate the three or four transfer gates per pixel by using advanced techniques of the three poly-silicon layer formation and the low temperature diffusion that is able to exactly control the impurity profiling. Therefore, the interline transfer CCD image sensor permitting to independently readout signal charges from all the pixels every field, the so-called progressive scanning could be realized and new applications of the device, such as digital still camera and image capturing for an image processing were rapidly developed.

The interline transfer CCD image sensor is highly sensitive to blue light compared with the frame transfer CCD mentioned later, because of no transfer gates on the light-sensitive area. But the device has two drawbacks: a low fill factor for incoming light and a relatively high level of smear. The fill factor as

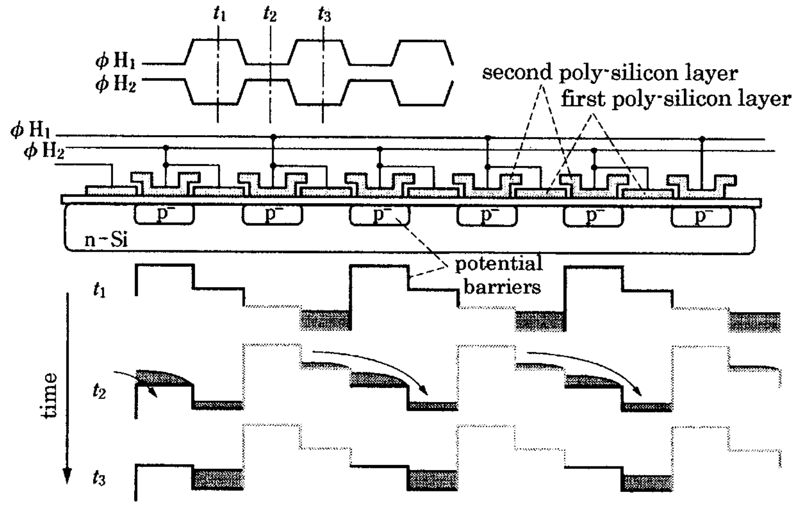


Figure C2.1.27. Cross-section of a horizontal shift register along the transport direction for two phase clocking.

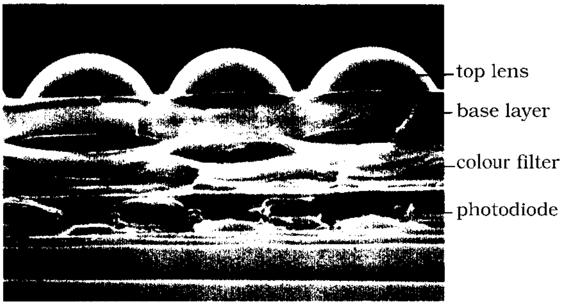


Figure C2.1.28. Cross-sectional view of micro-lens structure. (Photograph courtesy of H Komobuchi, Matsushita Elect. Co.)

measured by the ratio of active area to the total pixel area is usually below 30–40%, because the light insensitive vertical shift register occupies more than half of the imaging area.

To improve the low fill factor, on-chip micro-lens deposited on each pixel as shown in figure C2.1.28 were developed [19]. With the aid of the micro-lens array the effective aperture could be increased more than a factor of 2.5–3 and the smear could be also reduced by light convergence effect of the micro-lens.

As the interline transfer CCD imager is fitted to acquire both of the moving and resting optical images, the imager is now mostly used in a camcorder for both commercial and industrial purposes, and in a digital still camera. If restricting the purpose to the digital still camera application, a 2088(H) × 1550(V) imager of 3.45 μm × 3.45 μm pixels operated at a horizontal transfer frequency of 22.5 MHz and at a frame rate of 5 fps is on the market.

A new structured interline transfer CCD imager for improved resolution and light sensitivity was developed recently [20]. It is called a pixel interleaved array CCD (PIACCD) and has two advantages: reduction of light insensitive area in the pixel and increase of spatial resolving power. These features can be achieved by mutually sliding a row of the octagonal pixels by half pixel interval towards a horizontal direction as shown in figure C2.1.29 and by implementing a triangular sampling.

Full frame transfer CCD

In a full frame transfer CCD image sensor shown in figure C2.1.30, vertical shift registers serve as light-sensitive elements in a period of image sensing and as charge transfer channels in a period of signal charge detection.

When the photon is incident on the sensor, the signal electron charges are collected in the potential wells induced under the transfer electrodes, where positive clock pulse voltages are applied. After a certain integration period, the charge packets caught in the wells are transferred downwards by an appropriate clocking. The charge packets arrived at each end of the vertical registers are moved to the horizontal shift register one line by line. And then these are quickly transferred towards a device output through the horizontal register channel and are converted to video signal voltages. In this procedure, the parallel to serial transfer is achieved. After sweeping out the charge packets from the vertical shift registers, the integration of the charges is started again.

As the full frame transfer CCD sensor has a simple structure and it is relatively easy to fabricate the device having a large number of pixels, the sensor has been especially useful for an astronomical observation, a scientific measurement and others for a long time.

On the other hand, the applications of this sensor are restricted to relative low-speed imaging alone. The reason is that the integration of photo-generated charges and the transfer of the signal charges cannot be done simultaneously, owing to any temporary memory in the device. Moreover, to avoid a

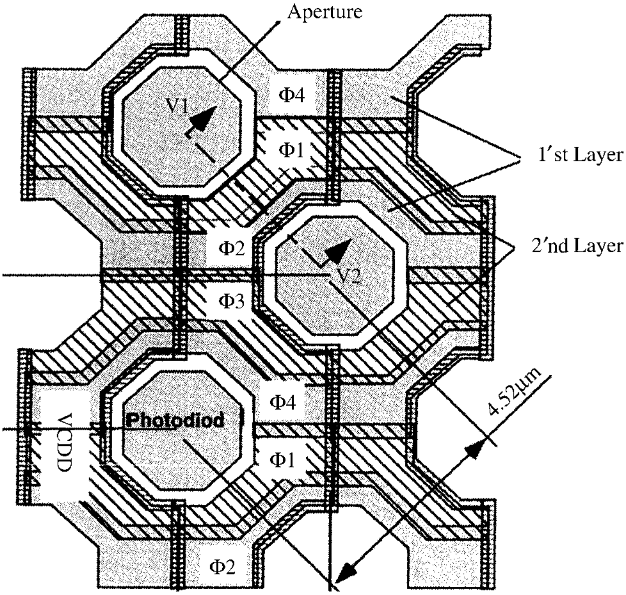

Aperture
V1
Φ4
Φ1
1′st Layer
Φ2
V2
2′nd Layer
Φ3
Φ4
Photodiod
VCDD
Φ1
4.52μm
Φ2

Figure C2.1.29. Schematic layout of pixel in a PIACCD showing the photodiode and the vertical shift register.

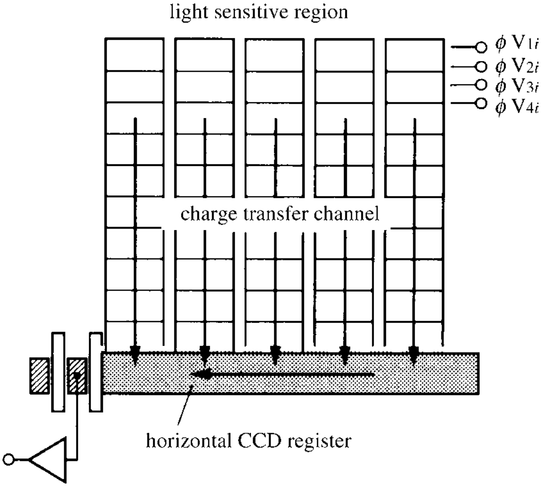

light sensitive region
ϕ V1i
ϕ V2i
ϕ V3i
ϕ V4i
charge transfer channel
horizontal CCD register

Figure C2.1.30. Organization of a full frame transfer CCD image sensor.

mixing of the transfer charges with the generated charges by the incoming light during the charge transfer period, it is required to shield the vertical shift registers from incoming light. For this purpose, a mechanical shutter is usually installed in front of the device.

At present, a 4096(H) × 4096(V) element frame transfer CCD image sensor with 9 μm × 9 μm pixels, driven at a horizontal transfer frequency of 10 MHz and a frame rate of 0.5 fps, respectively, has been available.

When a projected image on the full frame CCD sensor is moved along the vertical shift register and its speed can be synchronized with the velocity of a charge transfer in the register, it is possible to add up the charges that are generated by the same light signal in each register cell while transferring the charges. The operation calls time delay and integration (TDI) mode. As a signal multiplication proportional to a number of cells in the vertical shift register can be easily realized, the TDI is available for capturing of a moving object under low light level.

Frame transfer CCD

In a frame transfer CCD, each light-sensitive CCD register is extended by a CCD shift register of the same length, serving as a temporary memory. Charge transfer channels are formed by vertically dividing the photo-sensing and temporary memory sections as shown in figure C2.1.31. Vertical transfer electrodes are arranged on the charge transfer channels, perpendicularly to a channel direction. The horizontal CCD shift register transports the charge packets from the temporary memory into the output stage and then converts these into the successive video signal. Both of the temporary memory and horizontal CCD register are shielded from the incoming light by a metal film (Al or W–Si) through the appropriate insulating layer to avoid degradation of the signal due to the smear explained in section B5.3.2 in detail.

When the photon is incident on the light-sensitive area, the generated electron charges are collected in the potential wells induced under the transfer electrodes, where positive clock pulse voltages are applied in the same way as that of the full frame transfer CCD. After a certain integration period, the charge packets caught in each pixel are quickly transferred by the appropriate clocking and then stored

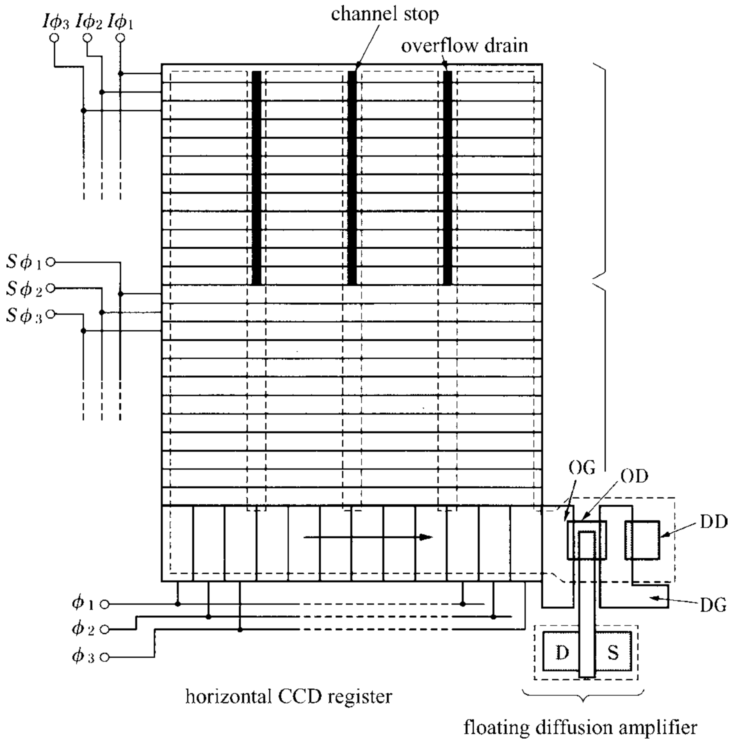


Figure C2.1.31. Organization of a frame transfer CCD image sensor.

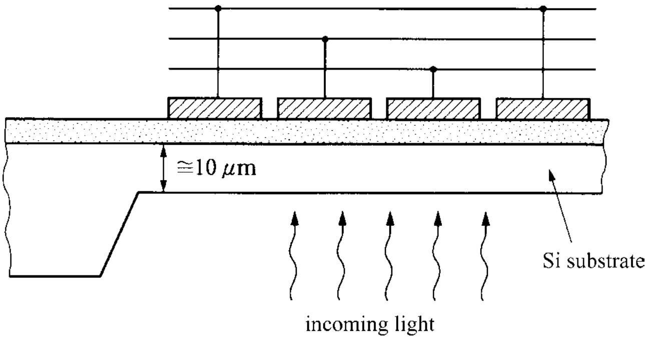


Figure C2.1.32. Back illuminated CCD image sensor.

in the temporary memory section. Once the charge transfer from the photo-sensing area to the temporary memory has been completed, the photon sensing and the charge integration are started again in the light-sensitive area, and the charge packets in the memory are shifted into the horizontal CCD register line by line. During the charge integration, the charge packets in the horizontal shift register are transferred towards a single output and are converted to a video signal voltage. This parallel to serial transfer is the same as described for the interline transfer CCD.

The first CCD image sensor was realized using the frame transfer CCD architecture that could be fabricated in a little simpler structure. But the sensor has the drawbacks of a large chip size and relatively large dark current, because the buried photodiode structure cannot be applied to photon sensing site. Moreover, the smear occurs easily during the charge transfer process from the imaging area to the temporary memory.

The sensitivity for blue light is usually low due to the light absorption of the semi-transparent electrodes covered on the light-sensitive area. To improve this problem, a back-illumination type image sensor as shown in figure C2.1.32 was developed. This sensor is being widely employed in various measurement camera applications, as it is highly sensitive over a wide spectral range from ultraviolet to infrared light.

A 4096(H) × 2048(V) element FT-CCD imaging sensor with 8.4 μm × 8.4 μm pixels driven at a horizontal transfer frequency of 37.125 MHz and at a frame rate of 60 fps has been developed [21].

In addition to these sensors, the electron-bombardment type CCD imager is also available. When a back-surface of the sensor is bombarded with high energetic electron particles corresponding to an input image, a large number of the secondary electrons can be generated in a substrate of the sensor and be caught into the pixels. This phenomenon is called the electron bombardment effect. This sensor is suitable for ultra low-light image sensing, for example medical and astronomy camera applications.

Frame–interline transfer CCD

To avoid the drawbacks of the frame and interline transfer CCDs, a frame–interline transfer CCD image sensor was developed. The device combines the light-sensitive section of the interline transfer CCD with the temporary memory of the frame transfer CCD as illustrated in figure C2.1.33. Both of the vertical-shift register located next to the pixels and temporary memory are shielded from incoming light in a similar manner to that of the previously described devices.

After the charge integration, the charge packet collected in each pixel is moved to the vertical shift register. Once the entire charge packet has been shifted to the vertical registers, the charge integration in the imaging section is started again. Meanwhile, the charge packets located in the vertical registers begin

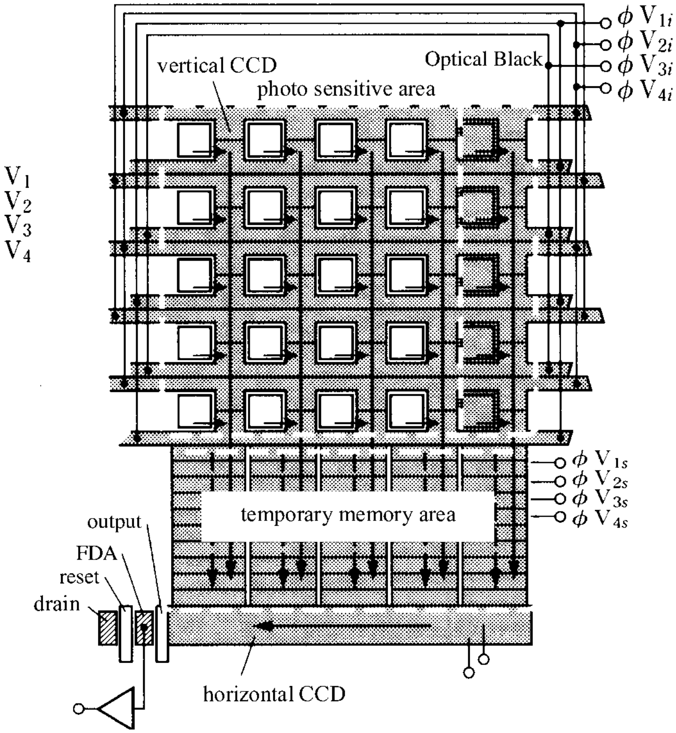

vertical CCD
Optical Black
photo sensitive area
ϕV_{1i}
ϕV_{2i}
ϕV_{3i}
ϕV_{4i}
V1
V2
V3
V4
ϕV_{1s}
ϕV_{2s}
ϕV_{3s}
ϕV_{4s}
temporary memory area
output
FDA
reset
drain
horizontal CCD

Figure C2.1.33. Organization of a frame-interline transfer CCD image sensor.

to transfer towards the temporary memory section. This transfer from the imaging section to the temporary memory has to be done as quickly as possible to avoid the smear that causes degradation of image signal. The degradation is inversely proportional to the vertical transfer speed. The higher the transfer pulse frequency brings about the smaller degradation of the signal as shown in figure C2.1.34 [22]. As the vertical shift register using low resistive poly-silicon electrode can be driven at the clock pulse frequency up to 1 MHz, the reduction in smear becomes possible in the frame–interline transfer CCD imager. The conversion of the parallel charge packets in the temporary memory section to the serial

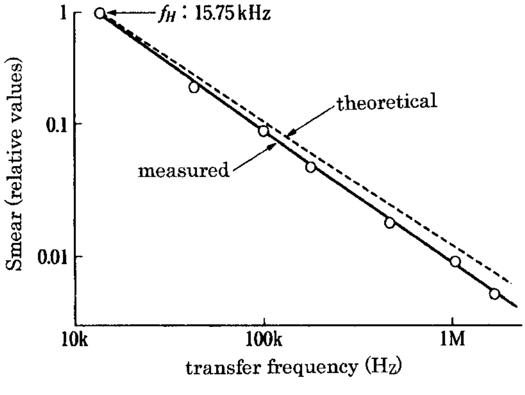


Figure C2.1.34. Effect of the charge transfer speed upon the signal degradation in a frame–interline transfer CCD.

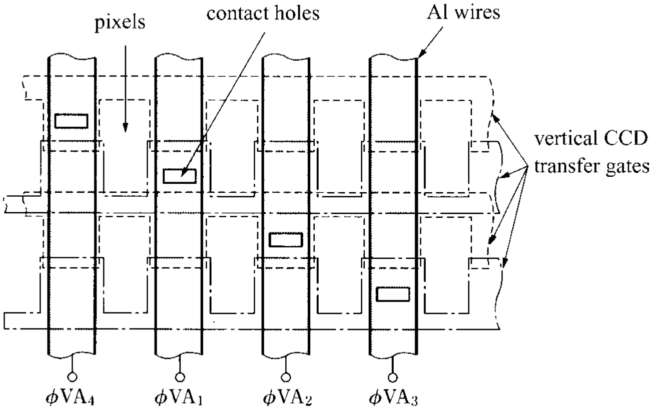


Figure C2.1.35. Structure of a poly-silicon film electrode backed with aluminium wires.

packets in the horizontal output register is analogous to that of the frame transfer CCD imager described previously.

Application of the frame–interline transfer CCD imager is restricted to the field where the acquisition of the high-quality image is requested, for example the broadcasting TV camera and others. The reason may be that the device is costly due to a relatively large chip size and a requirement of two kinds of the clocking pulse. One is required for the quick charge transfer from the imaging section to the temporary memory and other is for the slow charge transfer from the temporary memory to the output shift register, respectively.

A new structure of the transfer electrode backing the poly-silicon film electrode with aluminium wires as illustrated in figure C2.1.35 has been developed [22]. The sheet resistance of it is two orders lower than that of the poly-silicon. The new electrode is now employed in many frame–interline transfer CCD imagers for the broadcasting studio camera.

A 1936(H) × 1086(V) FIT-CCD imager of 5 μm × 5 μm pixels driven at a horizontal transfer frequency of 74 MHz and a frame rate of 30 fps is normally used in the field of advanced high-definition TV cameras [23].

C2.1.2.4 Device configuration of X–Y addressed MOS

An X–Y addressed image sensor (MOS X–Y) had a simple structure and could be fabricated in a somewhat simple technology. Nevertheless, there was a decisive drawback: quantities of kTC noise brought about from a large capacitance of the vertical signal line and a channel conductance of the horizontal select-switch.

By the recent progress of a fine process technology, it became possible to realize a CMOS X–Y imager with the same fill-factor or pixel sizes as that of the CCD, even though some integrated circuits together with a light-sensitive element are built within each pixel. Further, a CMOS active pixel image sensor (APS) having a small single-stage CCD fabricated in each pixel was realized in 1994. In the APS, a signal-to-noise performance equal to, or better than, that of the CCD imager will be expected in future.

The APS is shown schematically in figure C2.1.36. Operation of the device is as follows. During the charge integration, the photodiode is reverse biased, and the transfer gate TX, reset transistor R and the row selection transistor S are biased off. Following the charge integration, all pixels in the row to be read

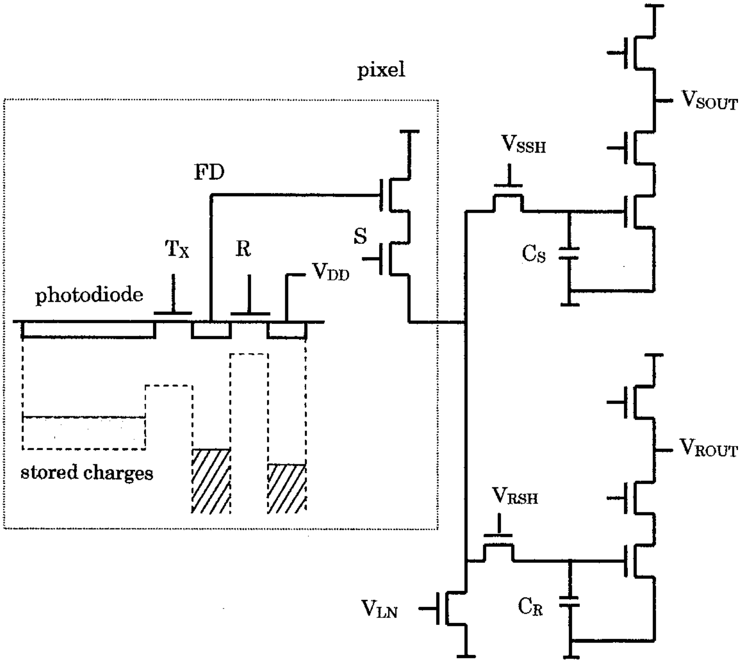


Figure C2.1.36. Schematic illustration of active pixel image sensor. Dotted line shows boundary of in-pixel circuits. The remainder of circuit is at bottom of column.

are addressed and are read out simultaneously onto the column lines by switching on the row selection transistor S. This activates the source follower output transistor in each pixel. Reset gate R is then briefly pulsed to reset the floating diffusion output node ED. The output of the source follower is then sampled into a holding capacitor CR at the bottom of the column. The transfer gate TX is then pulsed to transfer the signal charges stored in the photodiode to the floating diffusion output node ED. The new source follower output voltage is sampled onto a second holding capacitor CS at the bottom of the column. Storing the reset level and the signal level at the separated capacitors and detecting the difference of both levels permits correlated double sampling (CDS) which suppresses reset noise (kTC noise) from the floating diffusion node of the pixel and $1/f$ noise. And it also cancels pixel-to-pixel threshold variations of the source follower output transistor within the pixel. Most of these column fixed pattern noise-cancelling circuits are constructed by CDS circuits. However, the variation of the circuits is liable to generate column-wise fixed pattern noise and the column fixed pattern noise-cancelling circuits make the chip large.

A CMOS image sensor with a simple fixed pattern noise reduction technology has been reported [24]. This sensor is mainly constructed from a pixel array, I–V converter and CDS circuit. A circuit chain from the pixel to the I–V converter with a CDS is shown in figure C2.1.37.

The operation scheme of the sensor can be explained in aid of the operation timing illustrated in the same figure. A row of pixel circuits is selected in one horizontal scanning period by a row-select pulse Row(m). To obtain a reset level of the pixel signal, FD in the corresponding pixel is reset by a column reset-pulse Rst(n) just before the pixel period. Row(m) also enables the column readout pulse Col.-Rd(n) to be transferred to readout gate M2 through an X–Y addressing transistor M1. As Col.-Rd(n) rises at the centre of the pixel period, signal charges in a photodiode in the imaging area

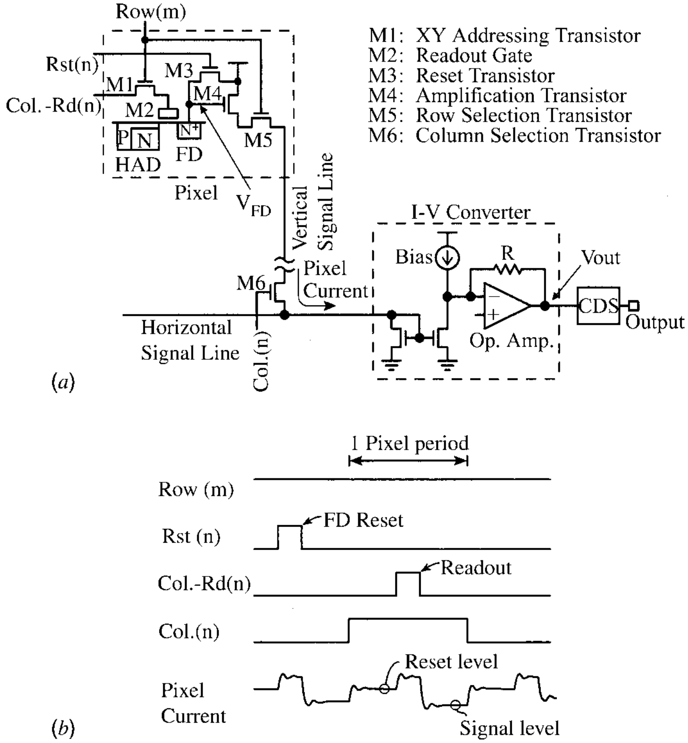

Row(m)
Rst(n)
Col.-Rd(n)
M1
M2
M3
M4
M5
P N
N+
HAD
FD
Pixel
V_FD
Vertical Signal Line
M1: XY Addressing Transistor
M2: Readout Gate
M3: Reset Transistor
M4: Amplification Transistor
M5: Row Selection Transistor
M6: Column Selection Transistor
I-V Converter
Bias
R
Vout
CDS
Output
Op. Amp.
M6
Pixel Current
Horizontal Signal Line
Col.(n)
(a)
1 Pixel period
Row (m)
Rst (n)
FD Reset
Col.-Rd(n)
Readout
Col.(n)
Pixel Current
Reset level
Signal level
(b)

Figure C2.1.37. (*a*) Circuit chain from pixel to CDS and (*b*) operation timing. M1: XY addressing transistor, M2: Readout gate, M3: Reset transistor, M4: Amplification transistor, M5: Row selection transistor, M6: Column selection transistor.

are transferred to FD. In the same pixel period, the column selection transistor M6 is switched on by the column selection pulse Col.(*n*). As a result, a selected pixel outputs a pixel signal current with a reset level and a signal level in a pixel period to the *I*–*V* converter through a vertical signal line and a horizontal signal line. The *I*–*V* converter receives the pixel signal in the current mode and converts it to the voltage mode. A CDS circuit subtracts the reset level from the signal level of a pixel signal to obtain the fixed pattern noise-free signal.

Furthermore, the CMOS image sensor is superior to any other imaging sensor in power consumption. Consequently, application to the mobile equipments will be especially anticipated.

A 1440(H) × 2160(V) array of 10.5 μm × 10.5 μm pixels operated at a horizontal scanning frequency of 12 MHz has been developed for the digital still camera [25]. Also a 2/3 CMOS imaging sensor for high-definition TV has been also reported lately [26].

The greatest drawback of the CMOS image sensor is poor signal-to-noise performance, because of the large dark current from the pixel. A significant reduction of the dark current by realizing the pinned photodiode structure is a pressing need of the hour.

Meanwhile, the CMOS image sensor has a distinctive feature that enables various sophisticated electronics to be integrated on the chip. In the beginning, integration of on-chip analog-to-digital converter is being explored. A 2357(H) × 1728(V) array of 7 μm × 7 μm pixels with on-chip 10 bit A/D converter and multi-readout ports, operated at a frame rate of 240 fps, is also being developed [27].

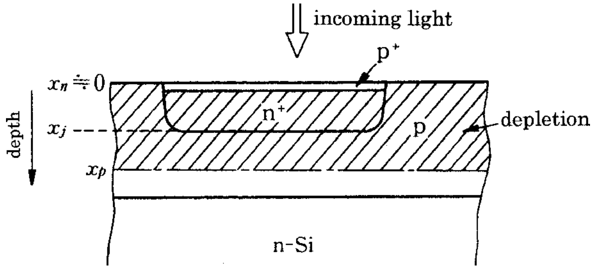


Figure C2.1.38. Cross-sectional view of the pinned photodiode.

C2.1.2.5 *Typical performance of area image sensors*

Sensitivity

Light-sensitive elements typically employed in the interline transfer CCD and the frame–interline transfer CCD image sensors are the pinned photodiode with n^+pn structure as shown in figure C2.1.38 [28]. The charge carriers photo-generated within the depletion region expanding into both sides of the n^+p junction are stored in a junction capacitance. Designing the pinned photodiode, it is important to precisely control both of the n-type dopant concentration and the n^+p junction depth x_j to fully deplete the n^+ layer, whenever the photodiode is reverse biased. A shallow layer of holes pinning the surface of the photodiode to the ground potential is, furthermore, provided on the surface to eliminate a recombination of the photo-generated carriers with surface states located near the surface. This is also useful for achieving a significant reduction of dark current.

Figure C2.1.39 shows a plot of the typical spectral responsivity of the pinned photodiode. It is found that the light sensitivity at a short wavelength region is hardly influenced by the hole inversion layer.

The photoelectric conversion characteristic that gives a relation between the incoming light on the imaging sensor and the signal output from it, is indicated by the following equation

$$i = kE^{-\gamma} \tag{C2.1.2}$$

where i is the signal output from the image sensor, E the quantity of incoming light and k a constant. As seen from the typical photoelectric conversion characteristic shown in figure C2.1.40, a slope of the plot

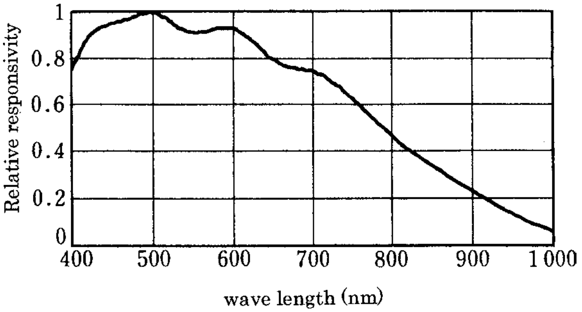


Figure C2.1.39. Spectral responsivity of the pinned photodiode.

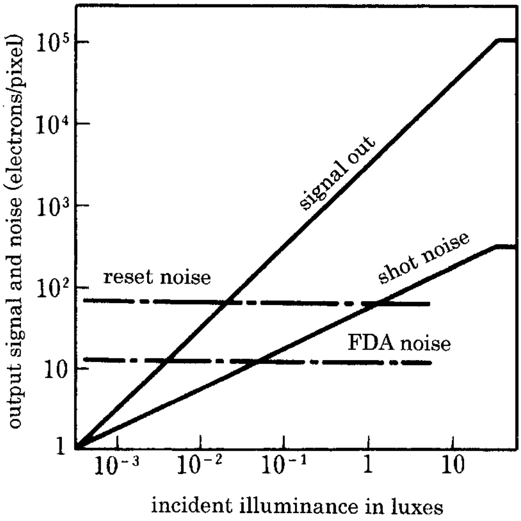


Figure C2.1.40. Photoelectric conversion characteristic of an interline transfer CCD image sensor.

is linear, that is $\gamma = 1$. The maximum of the signal output is usually restricted by a saturation level of the storage capacitance in the pixel. Meanwhile, the minimum is determined by either fixed pattern noise due to the variations of the dark current or random noise in the FDA. Dynamic range of the imaging sensor is defined as the ratio of its largest nonsaturating signal to the smallest detectable signal, i.e. the standard deviation of the noise under dark conditions. Since a sensor with higher dynamic can detect a wider range of scene illumination, it can produce images with greater detail. Today, the dynamic range of 70–80 dB has been usually achieved in a CCD imaging sensor on the market.

Dark current and noise

In the light-sensitive element, such as photodiode or MOS capacitor a little current keeps on flowing without incoming light. It is called the dark current. If the dark current is uniform through the pixels, the effect of the dark current on the signal output can be easily removed. However, when the dark current is varied from pixel to pixel, a fixed pattern, which looks like an image projected on a frosted glass arises. These variations of the dark current generally decide detectable low-light limit of the imager.

The dark current is increased with increase of temperature and is multiplied twice with every increase of 10°C. Therefore, measurements of the dark current and the fixed pattern noise are usually carried out by keeping a temperature of the sensor chip at 60°C.

It has been known that the dark current in silicon devices is a mostly generation–recombination (G–R) current due to the carriers thermally exited in a depletion region of the pn junction. Since the G–R current is dependent upon the number of interface state and/or surface state and their energies distributed in the energy gap of the semiconductor, the reduction of the interface and/or surface state density located near mid-gap energy is very important to decrease the G–R current. To achieve it, the pinned photodiode the surface of which is covered with a thin hole accumulation layer as shown in figure C2.1.38 has been employed in the CCD image sensor.

The noise generated in a solid-state imager can be classified into the fixed pattern noise (FPN) and the random noise. The former is usually observed as a spurious pattern fixed on the reproduced images. The main origin of the FPN may be considered due to both the variations of the effective light-sensitive area and of the dark current at every pixel.

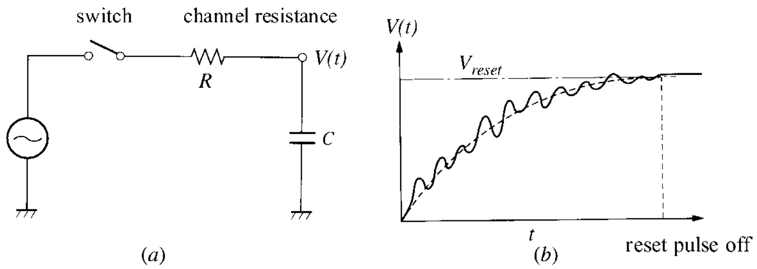


Figure C2.1.41. Noise equivalent circuit of the floating diffusion amplifier.

To decrease these variations, various improvements of the fabrication process, for concrete examples the accurate formation of the photo-resist films, the decreasing of the pitch error of the photomask and exact control of the light-shield film thickness, are being explored. When fabricating the CCD imager, an epitaxial wafer is generally used as a substrate, because a variation of relative resistivity along the wafer surface is less than 5%. In this case, the FPN tends to disappear particularly under high-light illumination levels.

Among the random noise in the CCD image sensor, the reset noise is commonly the most important. When the reset pulse voltage V_{rst} is applied to the reset gate in the FDA shown in figure C2.1.19, the floating diffusion is charged up and its potential is balanced at a certain voltage near the V_{rst}. But this reset voltage is fluctuated every reset procedure under the influence of the thermal noise caused in a channel resistance R as shown in figure C2.1.41 [29]. Assuming the channel resistance R and the floating diffusion related capacitance C, the fluctuation of the reset voltage, which is the reset noise, is represented using amount of the noise charge Q_{n} as follows:

$$\bar{Q}_n^2 = \mathrm{kTC}. \tag{C2.1.3}$$

The reset noise is otherwise known as kTC noise.

To eliminate the reset noise component in the signal, the CDS circuit is very helpful [30]. Figure C2.1.42 shows a typical circuit construction of the CDS, and schematically drawing waveforms of the signal and the sampling pulse. When the reset switch S_1 is pulsed at t_1, the potential on the output side of capacitance C_{C} is clamped to V_{CL}. Once the reset switch is pulsed off, the potential is changed in proportion to amount of the signal charge flowing into the floating diffusion. And then the reset switch S_2 is pulsed at t_2 and thus the potential is sampled on capacitance C_{SH} and is held. In this manner only the signal components can be extracted.

In the CCD image sensor with a lot of the dark current, the shot noise of the dark current becomes a dominant random noise and determines the lowest limit of a dynamic range.

Spurious signals

Blooming

Blooming is a phenomenon inherent to the solid-state image sensor. It is caused by illuminating some light-sensitive area over saturation. When lots of the charge carriers generated in strongly illuminated pixel overflow into the surrounding pixels, blurry bright spot spread out circularly around the pixel appears on a reproduced picture. If the carriers overflow into the vertical transfer channel, then a striped bright pattern comes out along it.

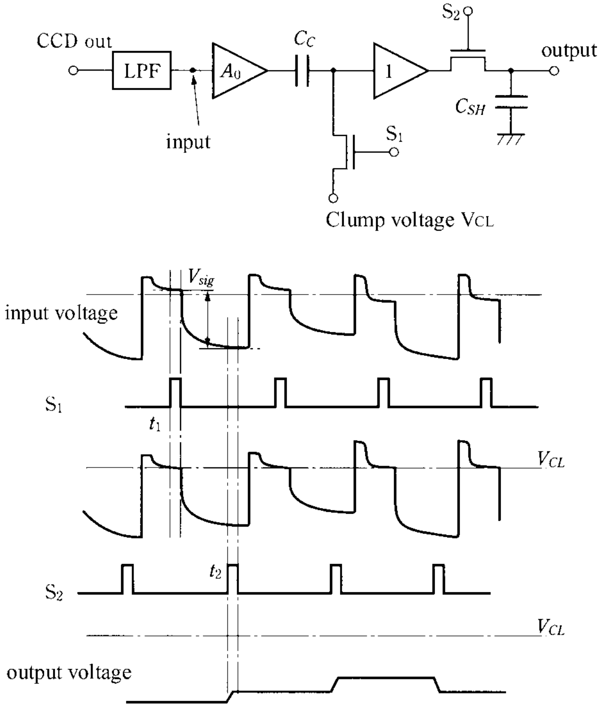


Figure C2.1.42. Correlated double sampling circuit and related input and output waveforms.

To suppress the blooming, an overflow drain was provided at a side of the photodiode [31]. The overflow drain serves to sweep out the excess charge carriers before the carriers overflow into the neighbouring photodiodes or the vertical transfer channels. The only drawback of the overflow drain is that an effective light-sensitive area in the pixel is cut down. Consequently, the vertical overflow drain (VOD) which serves to throw away the excess charge into a substrate has been extensively employed lately [32].

Operation of the VOD is as follows. The photodiode is formed in the p-well built on the n-substrate as shown in figure C2.1.43(*a*). The p-well must be entirely depleted by reverse biasing the p-well. If some pixel is strongly lilluminated, the potential of n^+ layer in the pixel drops and then the height of a potential barrier built up between the n^+ region and the p-well lowers as shown in figure C2.1.43(*b*). The excess charge carriers having higher energy than that of the barrier height among the charge carriers photo-generated in the pixel get over the barrier easily and flow out into the substrate. The CCD imager, which employs the VOD, enables the blooming to be suppressed below the practically imperceptible level.

Smear

The smear is a spurious signal that is observed like a blot blurred on the upper and lower sides of the image reproduced on the monitor as shown in figure C2.1.44, when some bright object is projected on the CCD image sensor. The phenomenon is inherent to the CCD image sensor and has three origins shown in figure C2.1.45. One is due to the leakage of a small amount of the incoming light into the vertical charge transfer register through the thin shield material covered on the register. The others are

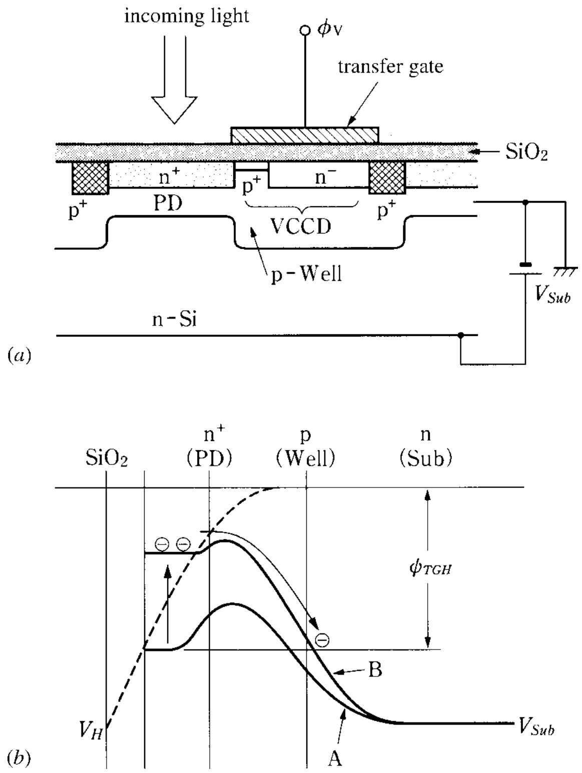

incoming light
ϕV
transfer gate
SiO2
n+
PD
p+
n−
VCCD
p+
p+
p-Well
n-Si
V_{Sub}
(a)
n+ (PD)
p (Well)
n (Sub)
SiO2
ϕ_{TGH}
B
A
V_H
V_{Sub}
(b)

Figure C2.1.43. (*a*) Cross-section of a vertical overflow drain. (*b*) Potential profile requirements for VOD: curve A, with empty well; curve B, with full well.

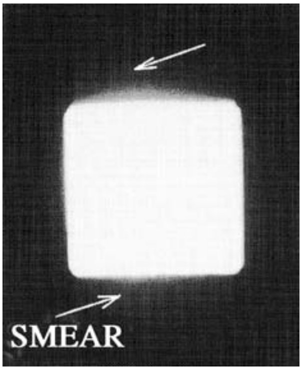


Figure C2.1.44. The smear reproduced on a monitor.

through the small gap left by an incomplete mask-alignment and through the thin oxide film between the gate electrode and substrate. In these cases, the smear appears even under lower illumination than the saturation level and thus remarkably deteriorates the reproduced image quality, especially when illuminating object is moved from side to side.

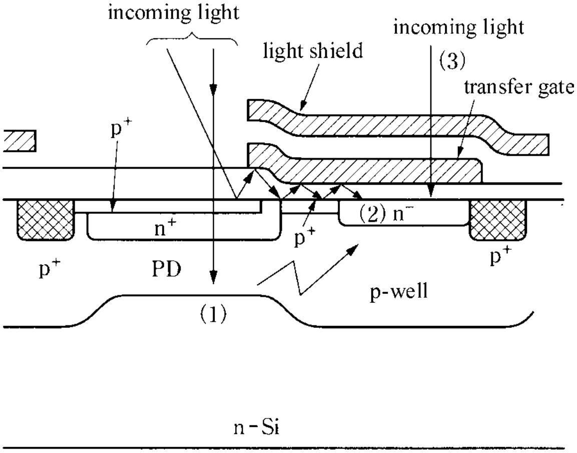


Figure C2.1.45. Generation mechanism of the smear.

As the smear level in the practical interline transfer CCD image sensor is now measured to be −80 to −100 dB and in the frame–interline transfer CCD is −120 to −140 dB, the influence of the smear can be practically ignored. To further suppress it, reduction in mask alignment error, improvement of lithographic process of an opaque thin film essential to shielding of the charge transfer channel and application of tungsten with greater optical absorption coefficient than that of aluminium as the shielding materials [33], etc are being explored.

Image lag

When all the signal charge carriers stored in each photodiode during a charge storage period cannot be readout entirely during a brief readout time, the image lag that shades off or trails a rear edge of the object rapidly in motion occurs.

As the signal charge carriers in the photodiode continue to be transferred into the vertical transfer register through the channel under the transfer gate, the amount of the charge carriers remaining in the photodiode decreases and the potential of the photodiode also drops. Once the potential reaches to a certain threshold, approximately given by the equation $V_G - V_S \leq V_T + n(kT/q)$, the channel conduction turns in a weak inversion. In the previous equation, V_G is the gate voltage, V_S the photodiode voltage and V_T the gate threshold voltage. Under this condition, transport of the charge carriers along the channel is due to diffusion process. Thus, the time necessary to sweep out the whole signal charge carries from the photodiode becomes longer and image lag occurs.

To solve this problem a new pixel structure described in section B5.1, which enables the diffused n-layer of the photodiode to be perfectly depleted under a given reverse-biased condition, was devised [34]. Using the structure, the signal charge carriers can be quickly passed through the channel, because the carrier transport mechanism in the channel remains to be drift. Accordingly, the whole signal charge carriers stored in the photodiode are completely swept out even though short readout duration and no image lag appear.

Resolution and Moiré

The spatial resolution of the solid-state image sensor is decided by the interval and shape of the light-sensitive elements in the pixel.

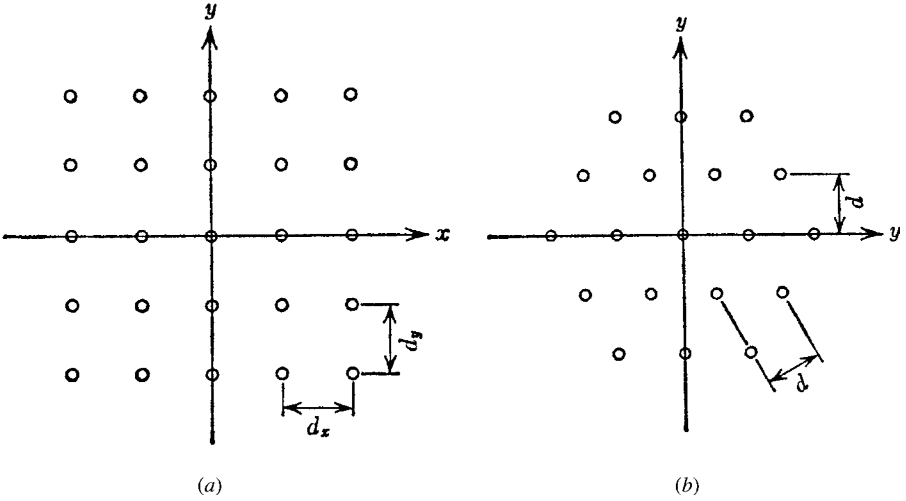


Figure C2.1.46. (*a*) Regular square lattice sampling structure. (*b*) Triangular lattice sampling structure.

The acquisition of image information by the image sensor is equivalent to looking at an original image through the tiny windows set on the points corresponding to the light-sensitive elements. Assuming a structure of the windows, that is sampling points, to be a regular square lattice illustrated in figure C2.1.46(*a*) and then calculating the two-dimensional Fourier transform of it, a rectangular reciprocal lattice pattern enclosed within limits of the Nyquist spatial frequency is obtained. The pattern illustrated on a spatial frequency domain is shown in figure C2.1.47(*a*). The meaning of this is as follows. The condition required for restoring exactly the original image from the sampled data alone is that the spatial frequency components contained in the original image have to exist only within this rectangular pattern. When a triangular lattice sampling structure as shown in figure C2.1.46(*b*) is applied, its Fourier transform is given by the hexagonal reciprocal lattice pattern as illustrated in figure C2.1.47(*b*).

Spectrum distribution of the special frequencies contained in common optical images may be considered to be almost uniform. Thus, this distribution is isotropic on the spatial frequency domain.

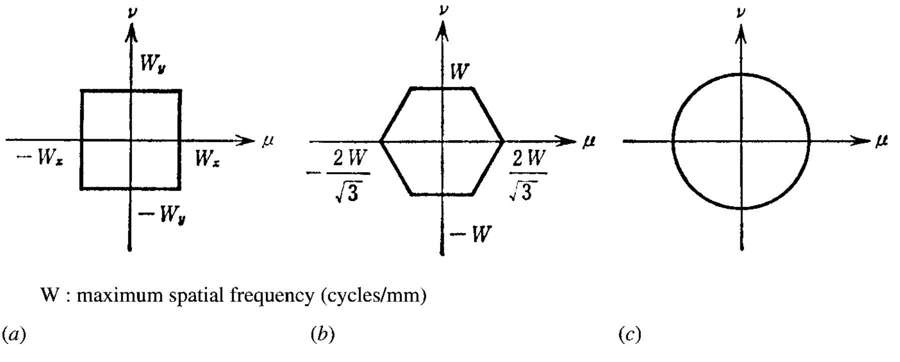


Figure C2.1.47. (*a*) Rectangular reciprocal lattice pattern. (*b*) Hexagonal reciprocal lattice pattern. (*c*) Isotropic spectrum distribution.

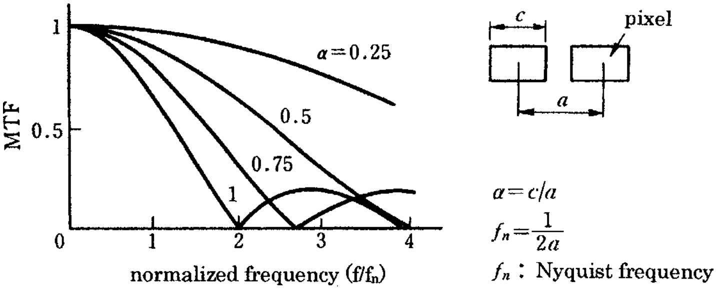


Figure C2.1.48. Size and interval of a sampling aperture, and related modulation transfer functions.

If reciprocal lattice patterns of both the square and triangular sampling structures are circumscribed to the circle where the spectrum components of a given image are confined as shown in figure C2.1.47(*c*), it is possible to perfectly restore the original state using either sampling structure. From the above-mentioned argument, it is found that a number of the pixels necessary to restore a given image using the triangular lattice sampling are about 15% smaller than that using the square lattice sampling.

In a practical sampling operation, the fact that the sampling point has a finite, though small, size in space cannot be ignored. Its size has an effect on the modulation transfer function (MTF) characteristics of the sensor.

Assuming that the shape of a sampling area, namely aperture function is square as shown in figure C2.1.48 and its response to the incoming light is uniform all over the area, the MTF is given by Fourier transforming the aperture function. The expression of the MTF in the horizontal direction is

$$\mathrm{MTF} = \frac{\sin(uc/2)}{uc/2} \qquad \text{(C2.1.4)}$$

where *c* is the size of the sampling area and *u* the spatial angular frequency. Hence, the MTF at Nyquist spatial angular frequency u_n is also given as follows

$$\mathrm{MTF} = \frac{\sin(\pi/2)(c/a)(u/u_\mathrm{n})}{(\pi/2)(c/a)(u/u_\mathrm{n})} \qquad \text{(C2.1.5)}$$

where *a* is the sampling interval. Sketches of various MTFs are also illustrated in figure C2.1.48. The smaller the sampling size is, the higher the resolution goes up. Light sensitivity of the device is, however, sacrificed, because decreasing of the sampling size is accompanied with smaller quantity of light incident on a sampling area.

There are two other factors that deteriorate the resolution. One is the diffusion of the photo-generated carriers into the neighbouring pixels [35] and the other is the incomplete transfer of the signal charge carriers in the CCD shift register [36].

To develop analytical approaches to the sampled-image acquisition system, a mathematical description of the sampling process must be investigated. To simplify the discussion about it, the sampling process using a linear image sensor alone is adopted.

For the sampling process, the 'samples' are elements of finite size whose amplitudes follow the incoming light variations of the space function $f(x)$ and are zero at all other areas. This type of sampling operation is shown graphically in figure C2.1.49.

The sampled data function $f^*(x)$ appears as an array of samples of finite size. The process may be thought of as being the result of multiplying a sampling function $s(x)$ and an incoming light image function $f(x)$ as shown in figure C2.1.49. This is a process of modulation, where a 'carrier' $s(x)$ is being

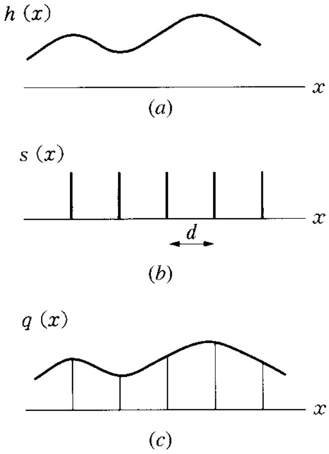


Figure C2.1.49. Finite aperture sampling operation.

modulated by an image function $f(x)$. Mathematically, the process may be represented by the following expression:

$$f^*(x) = f(x)s(x). \tag{C2.1.6}$$

In view of the fact that the sampling interval a is constant, thus making $s(x)$ a periodic function, it is possible to expand $s(x)$ into a Fourier series

$$s(x) = \sum_{n=-\infty}^{+\infty} C_n \exp\left(\frac{j2\pi nx}{a}\right) \tag{C2.1.7}$$

where the various C_n's are the Fourier coefficients of the exponential series. Substituting equation (C2.1.7) back in equation (C2.1.6), a representation of the sampling process is obtained.

$$f^*(x) = f(x)\sum_{n=-\infty}^{+\infty} C_n \exp\left(\frac{j2\pi nx}{a}\right). \tag{C2.1.8}$$

Carrying out the Fourier transformation of the sampled function $f^*(x)$, the following infinite summation is obtained

$$F^*(ju) = \sum_{n=-\infty}^{+\infty} C_n F\left[j\left(u - \frac{2\pi n}{a}\right)\right] \tag{C2.1.9}$$

where $F^*(ju)$ is the Fourier transform of the sampled sequence $f^*(x)$. This expression is very important in determining the effects of sampling on the information content of the original image $f(x)$. $F^*(ju)$ is seen to consist of a summation of weighted spectra, each of which is the same as the central term except for the weighting constant C_n and the shifted argument $j(u - 2\pi n/a)$.

The sampled spectrum $F^*(ju)$ is sketched in figure C2.1.50. The original spectrum $F(ju)$, from which the sampled signal spectrum $F^*(ju)$ is derived, is shown in this figure. The effect of the sampling process is to introduce a succession of spurious spectra which are proportional to the image signal spectrum and

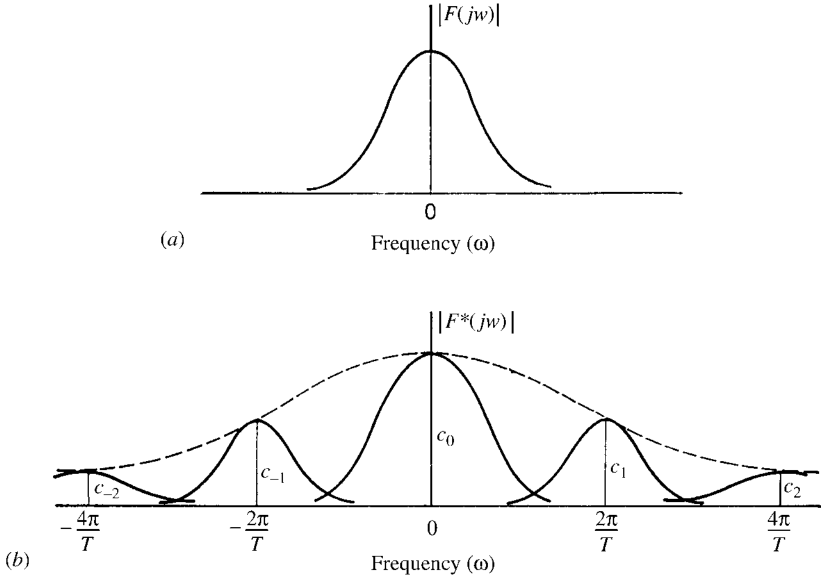


Figure C2.1.50. Spectrum of finite-aperture sampled function.

which are shifted periodically by a spatial frequency separation $2\pi na^{-1}$. It is evident that means to recover the original image information from the sampled signal must be found in any useful application.

The central signal spectrum can be extracted by rejecting the spurious higher spatial frequency spectra by means of an optical low-pass filter. However, it is evident that even if the filter is perfect, the spurious spectra components that enter into the central spectrum region can never be fully extracted, unless incoming image information contains no components above the Nyquist spatial frequency determined by the pixel interval. The spillover effect produced by an infinite spectrum such as that shown in figure C2.1.50 always produces distortion on the output of such a system. The distortion is called aliasing noise or Moiré. The aliasing noise generates the Moiré pattern at the region where the higher spatial frequency components are concentrated on, and remarkably spoils the quality of a reproduced image. Figure C2.1.51 represents the Moiré pattern observed when the image of a circular zone plate is reproduced on the monitor.

Image defects

When reproducing a video signal from the image sensor on a monitor, bright spots and lines, and dark spots and lines, etc are occasionally observed. These are called an image defect. Appearance of the image defects is the largest weak point of the imaging device, because these defects cause various obstacles and even a small defect is very striking. So the various efforts making an occurrence of the image defects minimum are being continued.

The origin of the image defect can be roughly divided into two categories: due to a quality of the wafer in use and due to a fabrication process of the device. The former has been mostly given rise from the striation of impurity concentration in the wafer. Both of the epitaxial wafer [37] and magnetic Czochralski wafer [38] enable the striation to be suppressed down below 0.5% recently.

Among the image defects arisen in a fabrication process, the bright spot is a typical one, especially striking in a dark scene. This is mainly caused from the crystal defects related to interstitial oxygen

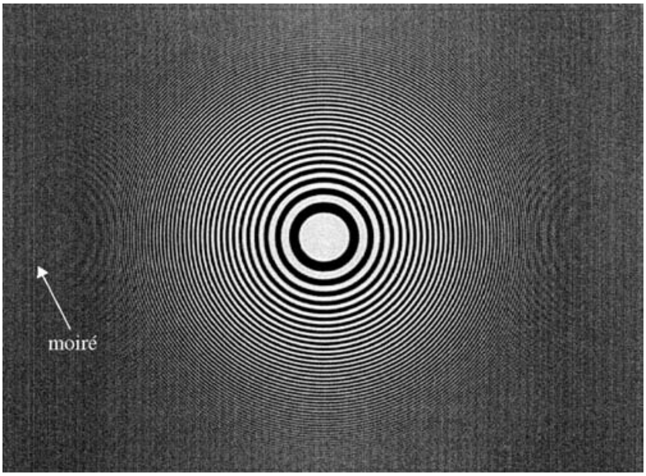


Figure C2.1.51. Moiré pattern arisen in capturing a circular zone plate.

atoms existing inside of the depleted layer. The bright spots can be reduced by using epitaxial wafer excluding oxygen.

An occurrence of the crystal defect is affected from various processing conditions, such as damage of the wafer surface during a dry etching process, compressive stress introduced around the LOCOS structure and Al wiring, and other various contaminations. When fabricating the image sensor, novel processes from which these defects hardly occur must be selected with the greatest possible care. Moreover, removal process of the contamination, such as an oxidation of the wafer in HCl atmosphere or a phosphor treatment has also been properly adopted according to circumstances. The intrinsic gettering, which enables the defects existing in an active layer of the wafer surface to be removed by diffusing high concentration impurities into the backside of the wafer or by making the defect intentionally absorbed into a deep region of the wafer, is also considered.

The dark lines and spots that are brought about from an imperfection of photo-lithography process such as misalignment of an etching pattern and nonuniformity of photo-resist film also deteriorate the quality of the reproduced image.

A coarse-grained pattern caused from pixel-to-pixel variations of the photo-sensing size is remarkable in a picture taken at a high-light level. In a normal photo-lithography process, influence of the light reflected at a surface of the light-shield layer on the irregularity of a photo-etched pattern cannot be ignored. Tungsten, well known as low reflective metal, is being adopted for the light-shield material.

C2.1.2.6 Conclusion

This section has described only a few aspects of CCD and CMOS image sensors, which are the basic operating principles, the various different architectures and some features of these devices, and some recent developments that improve the device's performance.

However, it has been impossible to cover all details of the technology and the performance of CCD and CMOS image sensors, because research and development works involving these sensors is extending to various fields.

At present, the technology of the CCD image sensor alone is mature among many image sensors and thus many reliable and low-priced products related to the CCD are on the market. Meanwhile, new technical trend that the MOS image sensor sadly defeated in struggle for light sensitivity against the CCD device in the past was resurrected as the CMOS image sensor serviceable to a digital still camera and smart sensing fields is strongly attractive.

History of the research and development work involving the CCD has been a succession of the technical challenges for improving both of the light sensitivity and the spatial resolution, since the CCD image sensor was born. This tendency will be unchanged in the future.

The pixel size of the image sensor was steadily cut down until now and came to be near 3 μm for the past few years. But the fact that the limitations of the pixel size exist is also sure. Its reasons are: realization of the optical lens systems having high resolving power gets into trouble, shot noise due to random character of the incoming light will be actualized with decrease of the charge carriers storable in the pixel and percentage of the light incident on the pixel may be lower according to make the pixel size approach the light wavelength.

The effort that increases the number of pixel elements is in progress in the field of a digital still camera, astronomy observation and defence applications. The possibility that a new application field of an ultrahigh resolution sensor appears on the market will be large in future.

Being accompanied with remarkable reduction of the noise level, the light sensitivity of the CCD image sensor has reached the level so high as even random character of the incoming light can be observed at the output of the sensor. The effort for elevating the light sensitivity over this level, however, stagnates now.

C2.1.3 CCD camera

A video camera converts a natural scene into an electrical video signal. For this purpose, a video camera consists of four major parts. Firstly, the optics system separates the tricolour components of an optical image of the scene and focuses them onto the surface of image sensors. Secondly, the image sensor converts the photons to electrons and scans the two-dimensional image to electrical video signal by scanning. Thirdly, the camera process circuits optimize video performance and make signals consistent with the external systems by processing the signals. Finally, other electronic features such as automatic exposure and automatic focusing.

C2.1.3.1 Optics

Optics overview

The purpose of optics in a CCD camera is to form an image of a subject on a light-detecting surface (image sensor). The optics system consists of a lens system and various components to perform functions such as colour-temperature conversion, optical low pass filter (LPF) processing and colour separation. Figure C2.1.52 shows the configuration of an optics system in a typical CCD camera. In this system, the camera first picks up a subject through the lens system. The received light then passes through a group of filters and a colour-separation mechanism, and the light of each colour is finally detected on respective CCD image sensors.

Lenses

Lenses have the role of forming a two-dimensional image of a subject. They make up the main section of a camera optics system and are designed according to specific ‘image sizes’. Table C2.1.5 lists the names assigned to various image sizes and their actual dimensions. A large image size means greater sensitivity or a higher S/N ratio on the one hand, and a larger and heavier lens system on the other hand. The brightness of an image is inversely proportional to the square of the *f*/stop.

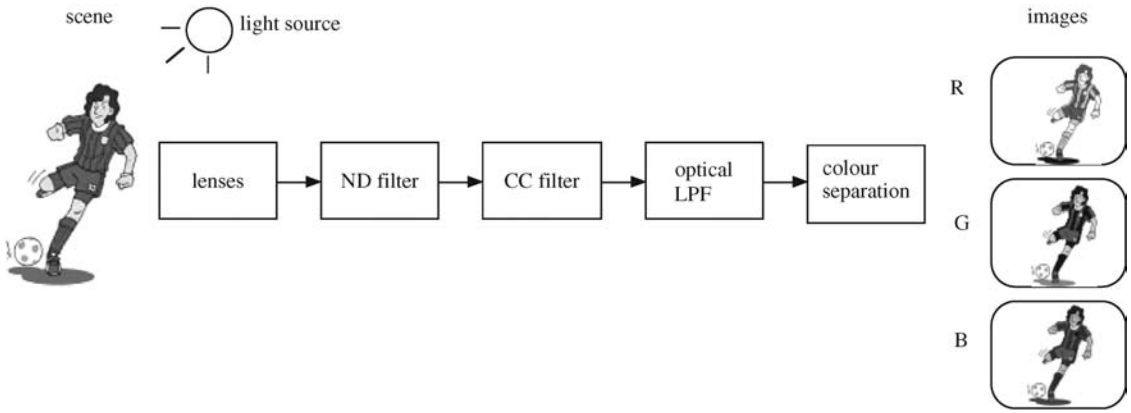


Figure C2.1.52. Camera optical system.

Table C2.1.5. Image sizes (mm) for 4:3 aspect ratio.

Format	*H*	*V*
1″	12.8	9.6
2/3″	8.8	6.6
1/2″	6.4	4.8
1/3″	4.8	3.6
1/4″	3.6	2.7

ND filter and CC filter

Both the ND (neutral density) filter and CC (colour conversion) filter make appropriate adjustments to the amount of light incident on the CCD image sensors. The ND filter attenuates incident light uniformly without relation to wavelength while the CC filter varies the amount of attenuation among R, G and B colours. In regard to the ND filter, it is also possible to adjust exposure by the camera's iris, but the value of the *f*/stop affects depth of field and the MTF. Exposure can also be changed by a CCD electronic shutter function, but this affects how motion is represented. The ND filter is used to vary exposure without changing the above characteristics.

A light source has its own particular spectrum characteristics. For example, outdoor daylight is bluish while light from an indoor tungsten lamp is reddish. The CC filter is the component that adjusts the spectrum characteristics of light incident on the camera. Adjustment of RGB signals can also be performed by electronic signals. However, to handle the dynamic range of the CCD efficiently, it is best to make the energy of these three colours as uniform as possible at their optoelectric conversion following colour separation. This is why a CC filter that operates in the optical domain is effective.

Optical LPF

Image capturing by a CCD means pixel-based sampling of a two-dimensional image. In sampling, if the input signal is not band-limited under the Nyquist frequency, a form of distortion called aliasing occurs. An optical LPF is used to prevent this from happening. A typical optical LPF makes use of crystal

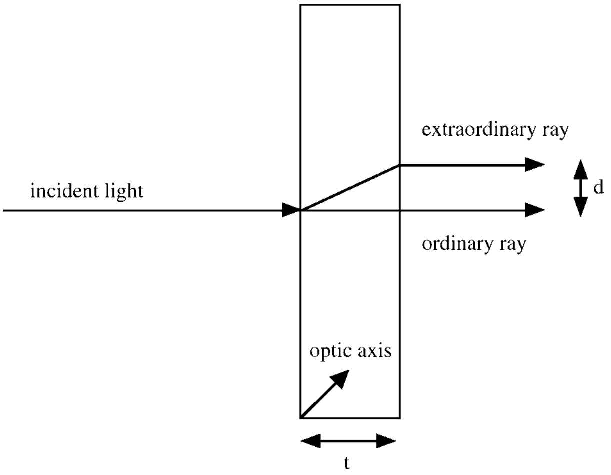


Figure C2.1.53. Birefringent optical LPF.

birefringence and achieves low-frequency passing by overlaying two separated images as shown in figure C2.1.53.

Separation d is expressed as follows:

$$d = \frac{N_o^2 - N_e^2}{2N_oN_e}t.$$

Here, N_o is the refractive index of an ordinary ray, N_e the refractive index of an extraordinary ray and t the thickness of an LPF.

This expression indicates that the value of d can be adjusted by changing the value of thickness t.

Frequency characteristics in this case are expressed as follows (figure C2.1.54).

$$g = |\cos(\pi f d)|.$$

Examining this figure, we see that the characteristics shown are not ideal. In general, t is selected so that g is null at the sampling frequency determined by the number of pixels. In this way, the intra-band modulation factor can be made high although aliasing still occurs to some extent.

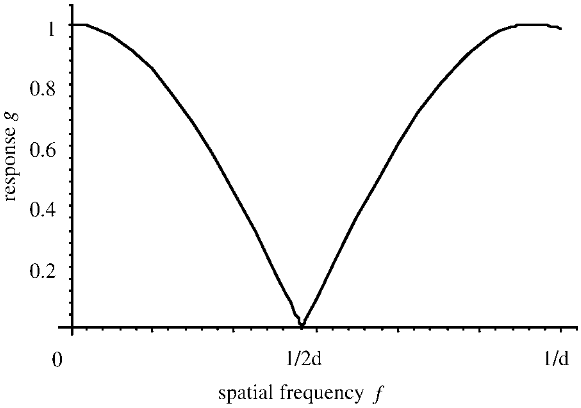


Figure C2.1.54. Characteristic of 1st order optical LPF, d: separation distance.

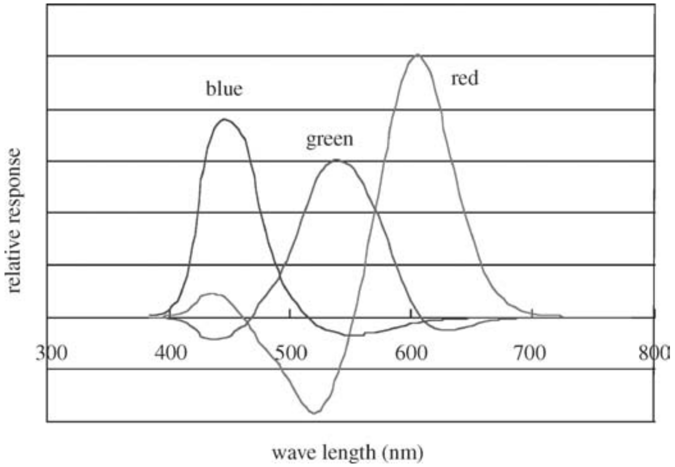


Figure C2.1.55. Camera taking curves.

Colour separation

A colour camera must capture three images each of a different colour to obtain colour information of the subject in question. These three colours are usually the primary colours: red, green and blue. The colorimetry of an electronic imaging system like television is determined by the display's primary colours and reference white. The spectral responses for each of the RGB camera channels, the camera taking curves must be equal to the colour matching function of the colour system of an electronic imaging system. The ideal camera taking curves for HDTV system is shown in figure C2.1.55. The negative values which the CCD imagers cannot output are produced by the camera signal processing called 'masking'.

Three images of different colours can be obtained by using either three image sensors or one image sensor. A camera using three image sensors employs a colour-separation optical system (prism) [39]. A colour-separation system that uses a prism features little loss of incident light with the resolution of each colour determined by respective image sensors, which means high-quality images. Television cameras for both broadcast and professional use employ this type of colour-separation system for this reason.

A camera using only one image sensor, on the other hand, employs a CCD in which a colour filter array (CFA) is placed over the image sensor. Single-imager colour pickup technology has been developed for consumer use. This approach is less expensive because it uses only one imager and makes optics simple, however, it suffers from problems including lower sensitivity and resolution, and false colour. Quite a few technologies have been developed to solve those problems.

A colour-separation prism has the function of breaking down incident light into the three primary colours of red, green and blue. Figure C2.1.56 shows a Philips type of colour-separation prism that combines selective reflection by a dichroic mirror and full reflection by the prism's surface to perform colour separation and double reversal (normal rotation). Trim filters placed between prism surfaces and each sensor bring the camera taking curves close to ideal ones.

A colour-separation system that employs a CFA uses one image sensor. It separates colours by placing a CFA as shown in figure C2.1.57 over the pixels of the image sensor, and obtains separate colour images for each type of pixel. Colour filters can be configured in a variety of ways. Figure C2.1.57(*a*) shows a Bayer pattern [40] in which four pixels make up one unit in a checkered arrangement. Two of these pixels are allocated to the green colour, which contributes most to the luminance signal,

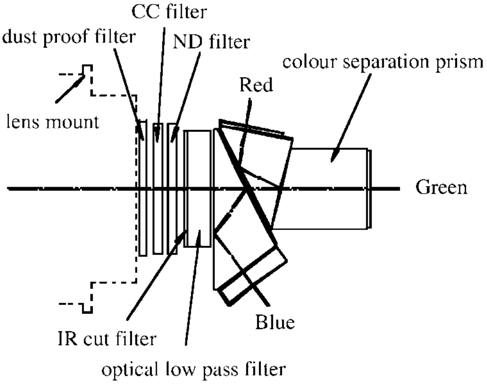

CC filter
dust proof filter
ND filter
colour separation prism
Red
lens mount
Green
Blue
IR cut filter
optical low pass filter

Figure C2.1.56. Colour separation optical system (courtesy of Fuji Photo Optical Co. Ltd).

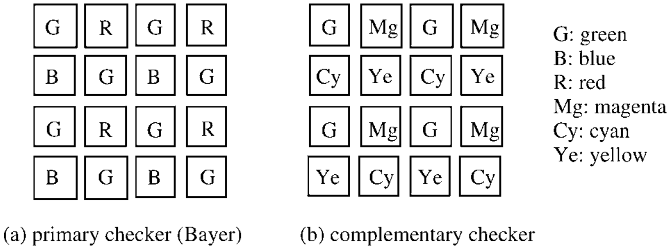

G R G R
B G B G
G R G R
B G B G
G Mg G Mg
Cy Ye Cy Ye
G Mg G Mg
Ye Cy Ye Cy
G: green
B: blue
R: red
Mg: magenta
Cy: cyan
Ye: yellow
(a) primary checker (Bayer)
(b) complementary checker

Figure C2.1.57. Colour filter array.

and one each of the other two pixels are allocated to red and blue. Figure C2.1.57(*b*) shows a complementary-checker pattern, which generates the line-sequential colour-difference signal in interlace scanning.

Figure C2.1.58 illustrates the colour-separation method of this colour-difference line-sequential system. Here, a filter is used to separate the signal into two frequency bands. The low-frequency component that sums two pixels becomes 2R + 3G + 2B and is nearly equal to the luminance signal. The high-frequency component obtained by taking the difference between two pixels, on the other hand, provides a G−2R or G−2B colour-difference component for each interlaced line.

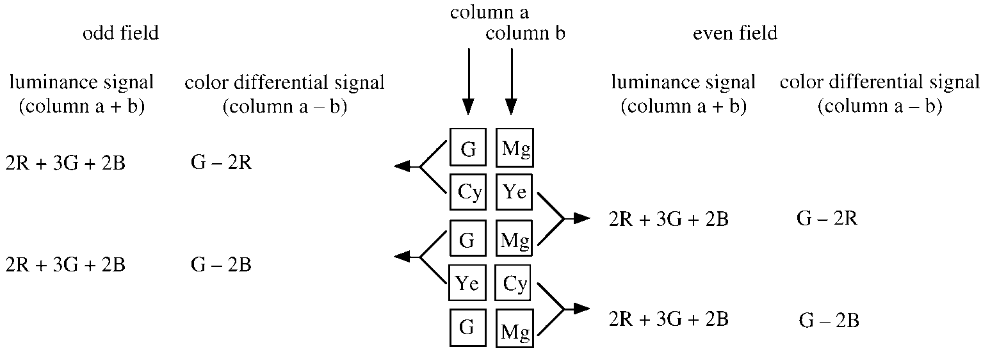

column a
column b
odd field
even field
luminance signal (column a + b)
color differential signal (column a – b)
luminance signal (column a + b)
color differential signal (column a – b)
2R + 3G + 2B
G – 2R
2R + 3G + 2B
G – 2B
G Mg
Cy Ye
G Mg
Ye Cy
G Mg
2R + 3G + 2B
G – 2R
2R + 3G + 2B
G – 2B

Figure C2.1.58. Ye–Cy–Mg–G field integration mode line alternated colour differential system.

Field sequential camera is a system that uses one image sensor and a rotating wheel of RGB colour filters to obtain a colour picture in a time-division manner. Shooting a fast-moving object with this system, however, results in colour breakup. A field sequential camera is therefore not appropriate for shooting such subjects.

C2.1.3.2 Drive circuit

Drive circuit overview

This section describes the drive mechanism for an area CCD image sensor. Figure C2.1.59 shows a diagram of an interline transfer (IT) type of CCD image sensor (see section C2.1.2.3). This type of sensor performs scanning by transferring within the sensor the charge generated by photoelectric conversion. The transferred charge is converted to voltage and output to the outside in the final stage of the scanning process. This operation requires electronic circuits driving vertical and horizontal transfer electrodes and supplying reset-gate pulses, bias voltages, etc.

CCD vertical and horizontal drive

Figures C2.1.60 and C2.1.61 show examples of pulse waveforms that are needed to drive an IT-type CCD. To begin with, the system sends charge stored in a photodiode out onto a vertical-transfer CCD (VCCD) register by opening a transfer gate that also acts as a vertical transfer electrode during the vertical blanking period. A vertical drive pulse transfers one step of charge every 1H during the effective period. Figure C2.1.60 illustrates a four-phase VCCD in which charge gathers underneath an electrode

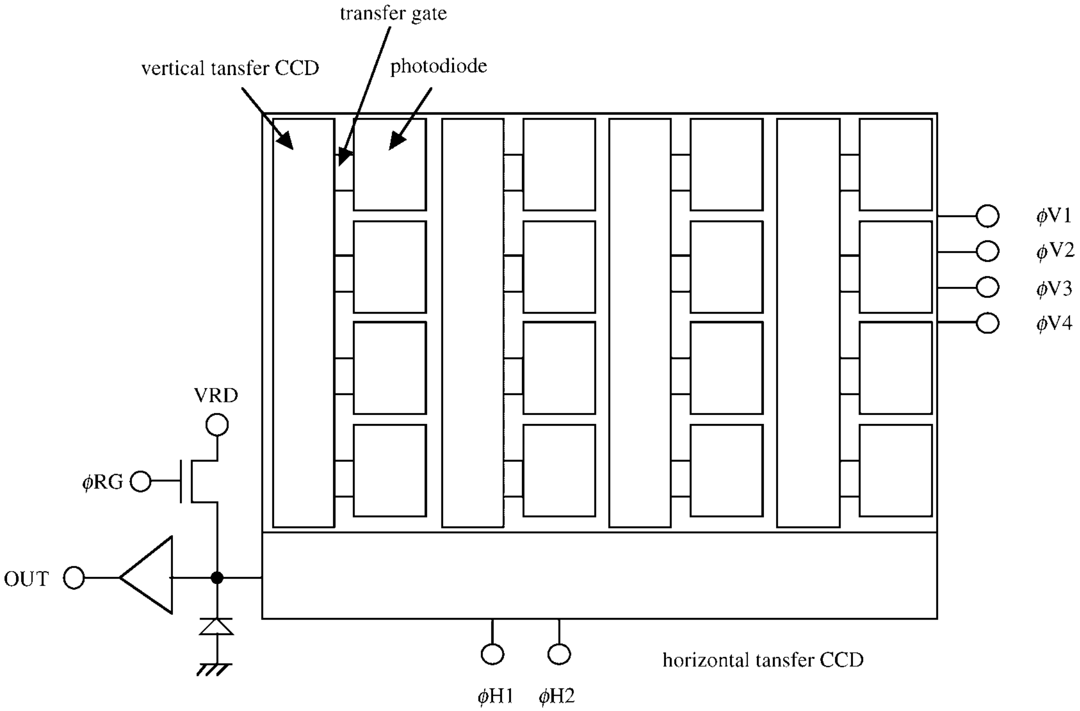


Figure C2.1.59. Configuration of Interline Transfer CCD.

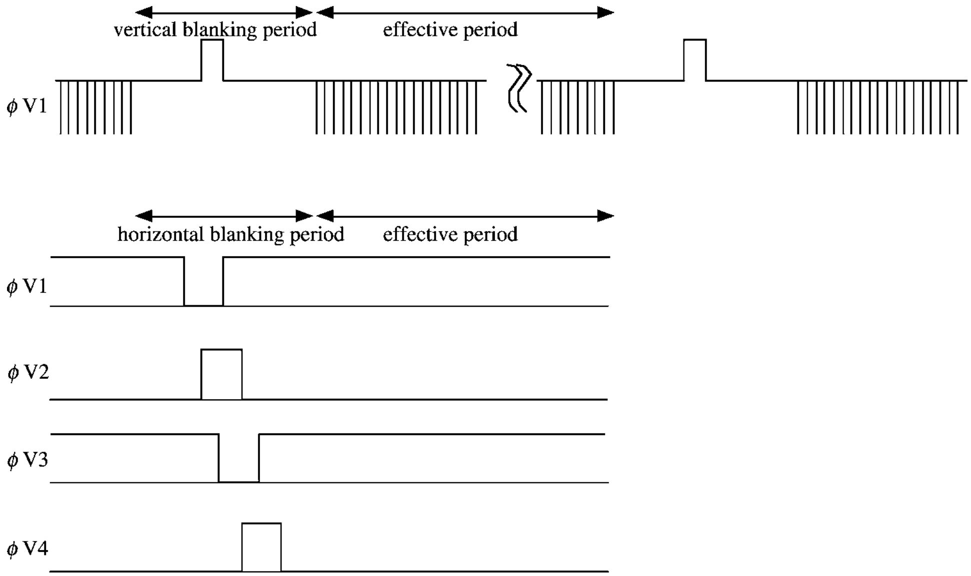

vertical blanking period
effective period
ϕV1
horizontal blanking period
effective period
ϕV1
ϕV2
ϕV3
ϕV4

Figure C2.1.60. VCCD drive pulses.

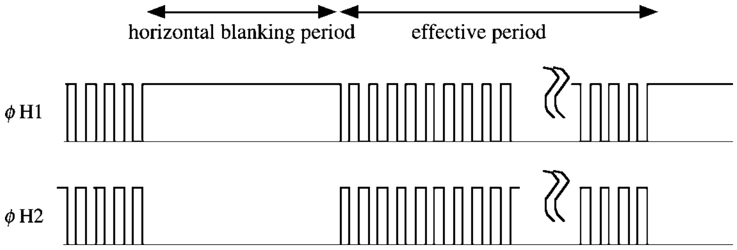

horizontal blanking period
effective period
ϕH1
ϕH2

Figure C2.1.61. HCCD drive pulses.

of low potential and becomes isolated from neighbouring pixels through a high-potential electrode. As shown in the figure, charge can be made to move in the direction from electrode V1 to electrode V4 by varying respective pulses. This constitutes one period during which one unit (pixel) of charge can be transmitted. Charge that has been carried by a VCCD is moved to a horizontal-transfer CCD (HCCD) during the horizontal blanking period.

The HCCD is often a two-phase CCD register with an implanted potential barrier. Figure C2.1.61 shows waveforms for this type of CCD. During the horizontal blanking period, charge is transferred from the VCCD to a location underneath the H1 electrode that goes high and is transferred to the output amplifier in the horizontal direction during the effective period. All charge is read out from an HCCD during a 1H effective period.

The system generates timing for these various pulses by a specific timing generator, general-purpose programmable logic devices (PLDs) and logic circuits. In addition, the system drives CCD electrodes by drivers of the form shown in figure C2.1.62.

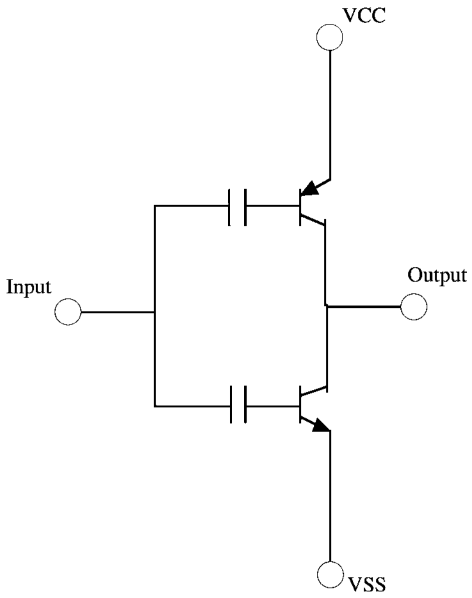


Figure C2.1.62. An example of a CCD drive circuit.

Gray-code counter

While not noise in the true sense, fixed-pattern noise can occur in the timing generator that generates CCD drive pulses as a result of frequency division by a counter. This noise can have negative effects such as nonuniform current due to the fact that a CCD handles extremely small signals. An effective means of preventing such fixed-pattern noise is to use a gray-code counter for which uniform consumption current can be maintained throughout the entire period. Figure C2.1.63 shows the difference in operation between a binary-code counter and a gray-code counter. The timing of a binary-code counter is such that many bits can make state transitions all at once. In a gray-code counter, however, the number of bits making state transitions is uniform across time.

Electronic shutter

An electronic shutter is used to control exposure time. It operates by treating the potential of the substrate as being higher than normal and disabling the barrier of the VOD so that charge stored in photodiodes can be discharged to the substrate [41]. An electronic shutter enables shutter speed in a digital still camera to be varied without a mechanical mechanism.

In addition to using automatic exposure control for movies, it can also be used for eliminating flicker from fluorescent lamps. Figure C2.1.64 shows how flicker can be eliminated when taking shots under 50-Hz fluorescent-lighting conditions with a camera having a 60-Hz field frequency. When not using an electronic shutter, the difference between the two frequencies generates a level difference every third field resulting in a 20-Hz flicker. If an electronic shutter lasting 1/100 of a second is now inserted, however, the same level will be produced across all fields thereby eliminating flicker.

Interlace scan

To output a video signal in interlace scanning, the positions of pixels to be read out must change every field. In an IT-type of CCD, this function also reduces the number of steps in the vertical transfer CCD

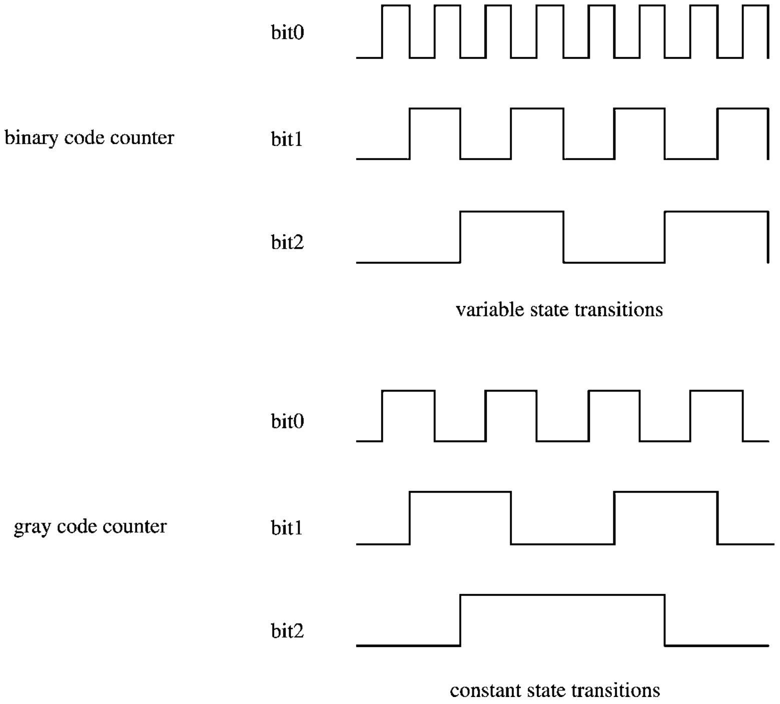

binary code counter
bit0
bit1
bit2
variable state transitions
gray code counter
bit0
bit1
bit2
constant state transitions

Figure C2.1.63. Binary code counter and gray code counter.

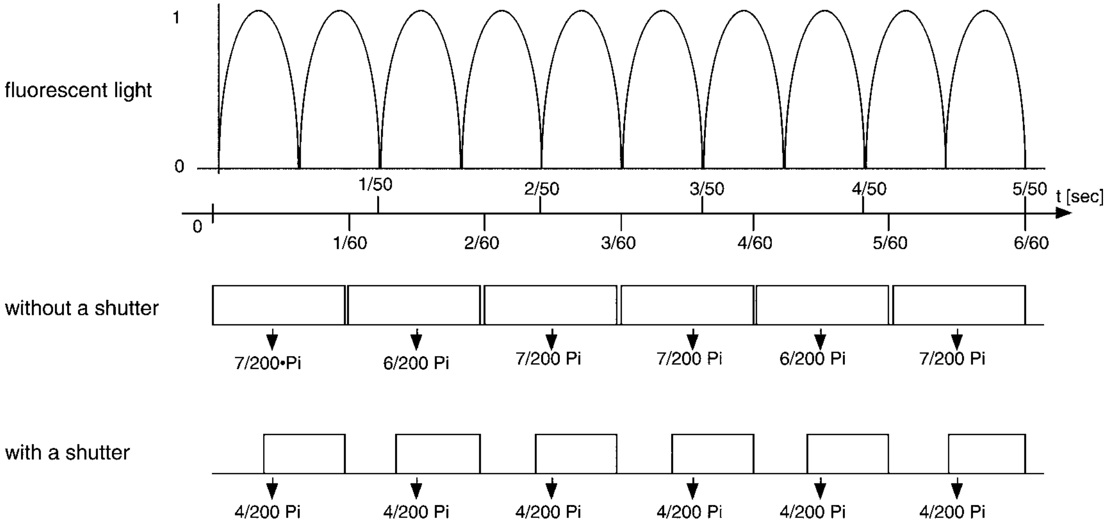

fluorescent light
1
0
1/50
2/50
3/50
4/50
5/50
t [sec]
0
1/60
2/60
3/60
4/60
5/60
6/60
without a shutter
7/200•Pi
6/200 Pi
7/200 Pi
7/200 Pi
6/200 Pi
7/200 Pi
with a shutter
4/200 Pi
4/200 Pi
4/200 Pi
4/200 Pi
4/200 Pi
4/200 Pi

Figure C2.1.64. Fluorescent light flicker suppression by electronic shutter.

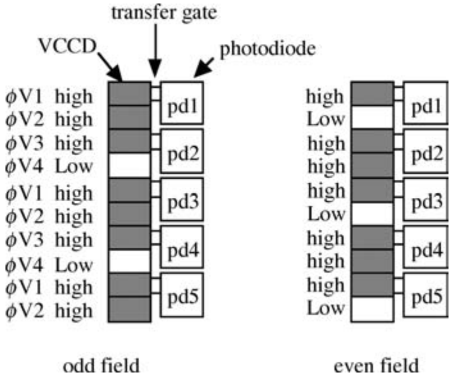

transfer gate
VCCD
photodiode
ϕV1 high
ϕV2 high
ϕV3 high
ϕV4 Low
ϕV1 high
ϕV2 high
ϕV3 high
ϕV4 Low
ϕV1 high
ϕV2 high
pd1
pd2
pd3
pd4
pd5
high
Low
high
high
high
Low
high
high
high
Low
odd field
even field

Figure C2.1.65. Interlace operation of CCD.

register. Figure C2.1.65 illustrates this concept. In progressive scanning, four electrodes are required per pixel, but in interlace scanning, four electrodes are sufficient for two pixels. Interlacing is performed as follows. For the odd field, ϕV4 in a four-phase CCD is made LOW and the transfer gate is turned ON so that signal charge of pixels becomes mixed in combinations of pd1–pd2 and pd3–pd4. For the odd field, ϕV2 is made LOW and signal charge becomes mixed in combinations of pd2–pd3 and pd4–pd5. The end result is interlace scanning.

C2.1.3.3 Camera signal processing

Signal processing overview

Although basic video signals are obtained by the photo-electronic conversion at the CCDs in a camera, the camera signal processing system has an important role in optimizing the video performance optimize and making signals consistent with the external systems. A CCD camera's amplification and signal-processing system comprises broadly a preamplifier and process amplifier. The main tasks of the preamplifier are to set CCD output to a required level and suppress noise, while the role of the process amplifier is to optimize the video signal through the correction and compensation processes described in this section.

Note that the above is a conceptual description of these two amplifiers. In actual hardware, they may be integrated as one unit depending on the type of camera.

Preamplifier

The preamplifier performs DC restoration, prepares a video signal of required voltage, and passes the result on to the process amplifier. This process includes CCD noise reduction. Figure C2.1.66 shows a typical preamplifier configuration. A noise reduction is performed at the CDS circuit after the first stage amplification. A several times wider bandwidth than the pixel read-out rate is required to secure a

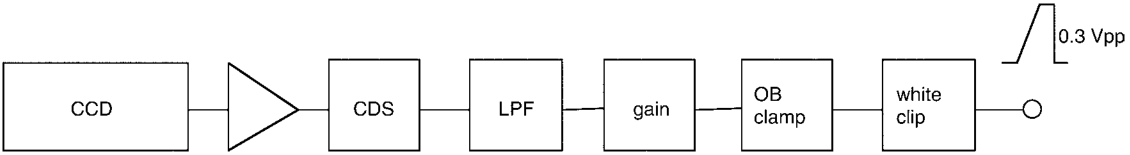

CCD
CDS
LPF
gain
OB clamp
white clip
0.3 Vpp

Figure C2.1.66. Configuration of a pre-amplifier.

sufficient noise reduction performance until this stage. Then the bandwidth is limited to the required video bandwidth by an LPF. The gain part amplifies the signals of RGB channels to make them consistent with the required input form of process amplifier and accomplish white balancing. The required signal voltage is taken to be 0.3 Vpp in this example. The OB clamp part performs the DC restoration of the signal by using the signals from the optically shaded part of the CCD as reference. The signal is passed from the preamplifier to the process amplifier after white clipping.

Correlated double sampling

The CCD output signal includes reset noise and amplifier noise originating in the on-chip readout amplifier. This noise, however, can be removed by signal processing [42]. The following describes CDS, the most common method for removing noise. Reset noise is dispersion in each pixel caused by the resetting of floating diffusion capacitance via FET channel resistance. This dispersion is fixed within a pixel. Accordingly, the circuit of figure C2.1.67 can be used to eliminate reset noise by taking the difference between the feed-through period and signal period for each pixel as shown in figure C2.1.68. At this time, the low-frequency components of random noise can also be reduced, which means that the circuit is also effective in reducing $1/f$ noise generated in the readout amplifier. Because this noise-removal circuit requires a frequency band that significantly exceeds the pixel clock frequency, it should be located at an early stage, the earlier the better.

DC restoration

Video-signal processing and reproduction require that DC restore be performed. The preamplifier uses the black level output from the CCD as a reference for this purpose. Here, black level output corresponds to the output coming from pixels shielded from light. Dark current is generated at every pixel of CCD and is added to the signal equally both for sections of the CCD sensor receiving light and shielded sections. For this reason, pixel output corresponding to only dark current can be used as a reference for fixing the black level (optical black (OB) clamp) enabling DC to be restored, as shown in figure C2.1.69.

White clipping

In the case of incident highlighting, the difference in saturation levels among the R, G and B CCD sensors results in white colouring. White clipping in a preamplifier is performed to prevent it. Specifically, the colour channel that saturates the quickest (generally green) can be used as a basis for clipping the signals of the other channels.

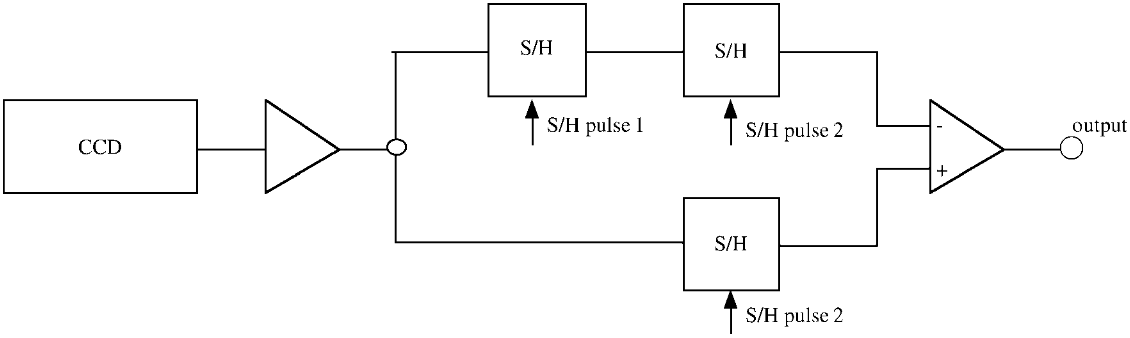


Figure C2.1.67. CDS circuit.

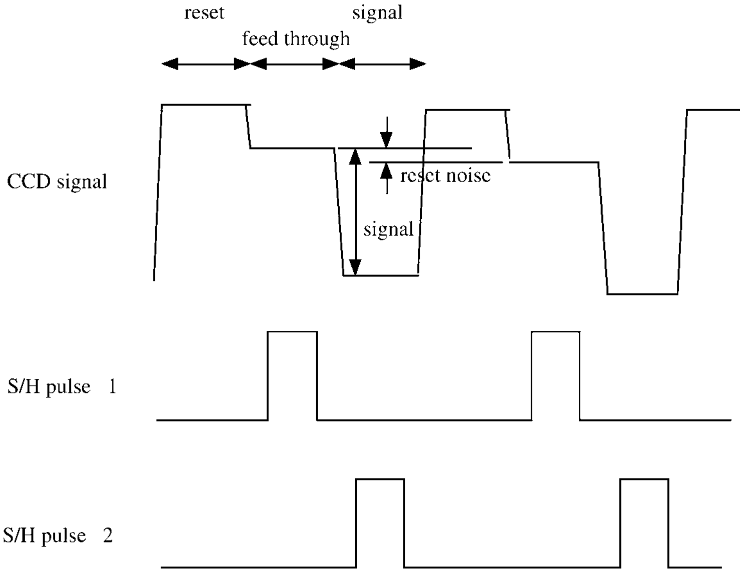

reset
feed through
signal
CCD signal
reset noise
signal
S/H pulse 1
S/H pulse 2

Figure C2.1.68. CDS operation.

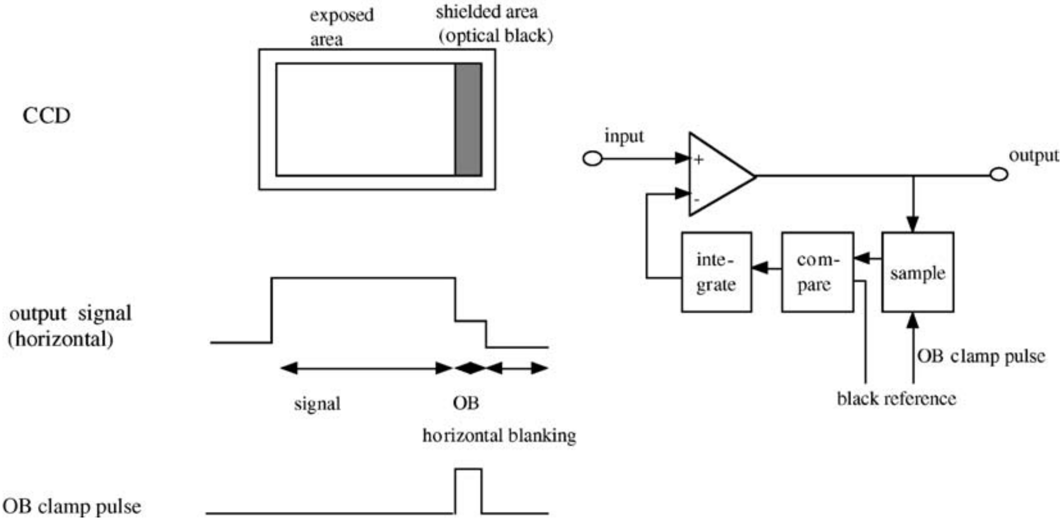

CCD
exposed area
shielded area (optical black)
output signal (horizontal)
signal
OB
horizontal blanking
OB clamp pulse
input
+
-
output
inte-grate
com-pare
sample
OB clamp pulse
black reference

Figure C2.1.69. OB clamp.

Process amplifier

The process amplifier performs various types of correction and compensation processing with respect to preamplifier output in order to make a signal consistent with the standard in question (e.g. NTSC, PAL and HDTV) [43–45]. Figure C2.1.70 shows a typical process amplifier configuration.

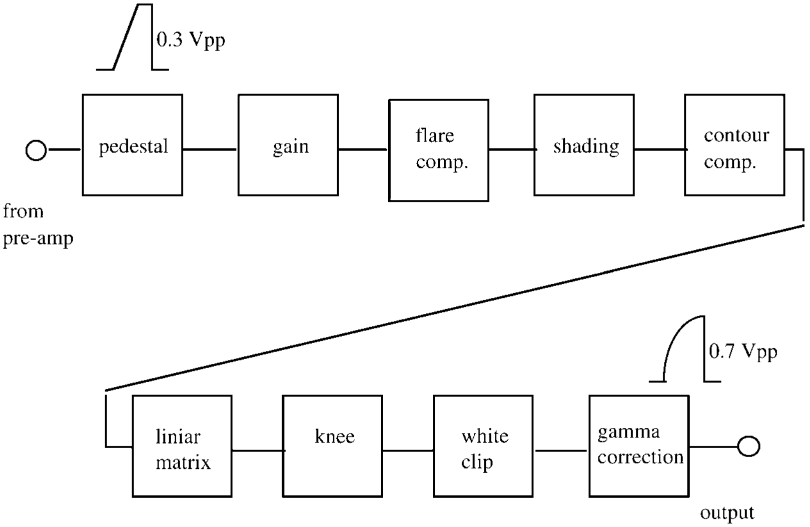


Figure C2.1.70. Configuration of a process amplifier.

Flare compensation

Flare is a phenomenon in which the CCD output-video-signal level rises as a result of stray light caused by light scattering in the optical system and at the surface of the CCD image sensor. Flare correction is performed by first determining the amount of flare from the video DC component and then subtracting that amount.

Shading correction

This circuit corrects for nonuniformity in black shading (offset) and white shading (gain) within the screen. Such nonuniformity originates in the lens system, colour separation system, and drive system and the CCDs. Black shading is corrected by subtracting the corresponding correction signal obtained by capping the camera. White shading is corrected by modulating the gain in accordance with the inverse of the sensitivity characteristics.

Contour compensation

The function of contour compensation is to emphasize contours, i.e. high-frequency elements, in a picture to improve sharpness. Basically speaking, it aims to compensate for high-frequency degradation due, for example, to the optical transfer function of the optical system, the aperture effect of the image sensor and the signal transfer characteristics of the electronic circuitry.

A contour compensation section can be configured in a digital signal processing circuit as shown in figure C2.1.71. Here, the function divides contour extraction into the horizontal and vertical directions: an n-tap transversal-type of high pass filter (HPF) is used in the horizontal direction while an m-tap HPF of the same type is used in the vertical direction. Boost frequency can be varied here by changing coefficients. The nonlinear processing section shown in the figure performs various processes such as coring, which clips a small-amplitude contour-compensating signal to reduce noise, and level-dependent processing, which minimizes a contour-compensated signal in dark sections.

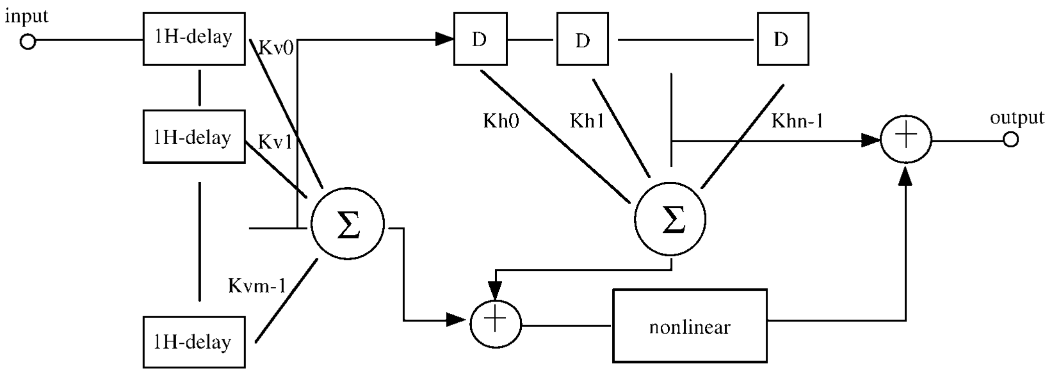


Figure C2.1.71. Contour compensation.

Gamma correction

A CRT receiver has nonlinearity in the light output versus input voltage characteristic that is called gamma characteristic. In current television systems, the camera side compensates for gamma characteristic to minimize the cost of receivers. The circuit that performs this function is called the 'gamma correction circuit'.

A CRT has a transfer characteristic expressed as

$$l = v^{\gamma_{\mathrm{CRT}}}$$

where l is the brightness of the CRT and v the input voltage. The camera gamma correction circuit compensates the input signal as

$$v_{\mathrm{out}} = v_{\mathrm{in}}^{\gamma_{\mathrm{CAM}}}$$

where v_{in} and v_{out} are the input and output voltage of the gamma correction circuit. γ_{CAM} is usually set to be 0.45, because a typical value of γ_{CRT} is 2.2.

In digital signal processing, gamma processing may be performed by a look up table (LUT) system or an arithmetical system.

Knee compression

This function compresses excess highlights by reducing the gain of the transfer characteristic above a certain level to make the highlights fit the system dynamic range. This improves the reproducibility of highlighted sections without losing all the information by white clipping.

Black and white clipping

The level range of a video signal is determined by the limitations of the system dynamic range including transmission path and other factors. It spreads from the minimum to the maximum digital value determined by the number of bits per sample in a digital system. Black clipping and white clipping are processes performed to keep the signal in that range.

Transfer characteristics

Figure C2.1.72 shows transfer characteristics in the process amplifier when performing gamma correction, knee compression, white clipping and black clipping. Here, gamma characteristics and digital

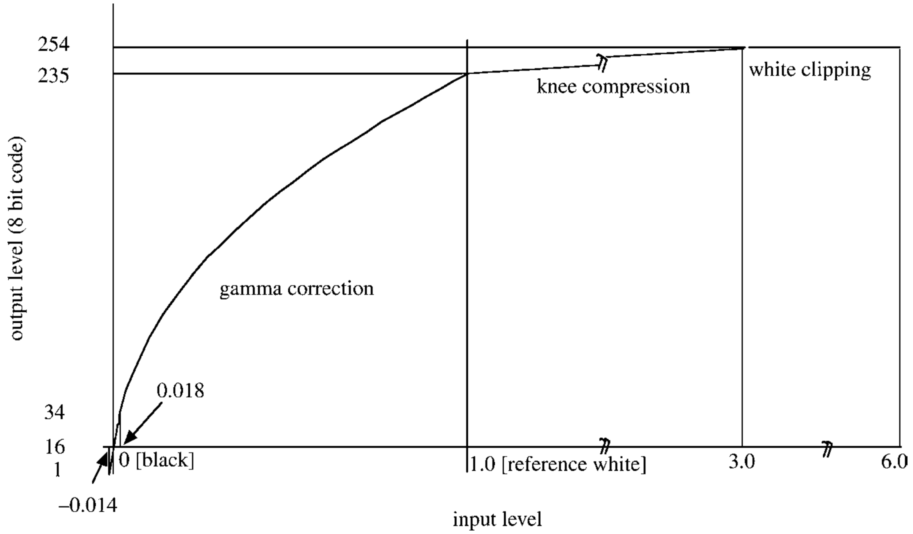


Figure C2.1.72. Transfer characteristics.

code conform to ITU-R Rec. 709. Knee point and knee slope are set to 100% and about 1/23, respectively.

Colour correction

In an electronic imaging system, the colour matching function determined from the chromaticity points of the three primary colours in the colour system becomes the spectral characteristics of the three primary colours in the pickup equipment. In general, these spectral characteristics include a negative portion, which cannot, however, be achieved in the colour separation optics of an actual camera. This is corrected in the electronic domain by a technique called masking process that is implemented in an analog or digital circuit by an operation expressed by the following expression

$$\begin{bmatrix} R_{\mathrm{out}} \\ G_{\mathrm{out}} \\ B_{\mathrm{out}} \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} R_{\mathrm{in}} \\ G_{\mathrm{in}} \\ B_{\mathrm{in}} \end{bmatrix}.$$

White balance

In a colour imaging system, colours are expressed on the basis of the three primary colours. Accordingly, white occurs on the display equipment when the levels of these three colours are equal. This, in turn, means that the signal levels of these three colours must be equal when the subject being shot by the camera is a ‘white object’ (an object that appears to be white when looked at by humans). The function for achieving this is called ‘white balance’, which involves a rough adjustment at the CC filter described earlier and a subsequent electronic adjustment of RGB signal levels.

Digitalization

Due to the fact that bit width and speed of analog-to-digital converters and digital signal processing ICs have reached a sufficient level, nearly all new cameras adopt digital signal process circuits for its camera process to take advantage of digital technology that enables extremely sophisticated signal processing including described earlier at less expensive cost.

C2.1.3.4 Automatic operation

Automatic operation overview

Cameras, especially those for the consumer market, incorporate a number of automatic functions like automatic exposure and automatic focus to simplify shooting. Digitalization has made it easier to incorporate complex and sophisticated functions.

Automatic exposure

Automatic exposure automatically keeps the brightness of the subject within the limited dynamic range of the camera. The system consists of the detection part, exposure control part and feedback loop. Figure C2.1.73 shows an example of the basic configuration of this function. Here, the lens iris and/or an electronic shutter are used to adjust the amount of exposure. The amount of adjustment is determined by specifying a picture area and extracting high-level portions so as to achieve a prescribed level. Note that the subject's field of depth changes when adjusting exposure by the iris and that motion resolution changes when adjusting exposure by an electronic shutter.

Automatic focus control

In general, an automatic focus function consists of focus or distance detecting part, focus ring actuator and loop connection. Information for focusing can be obtained from either the picture itself or by some other means. Figure C2.1.74 shows the former scheme in which feedback control sets the lens focus position so that high-frequency components of the image become maximum.

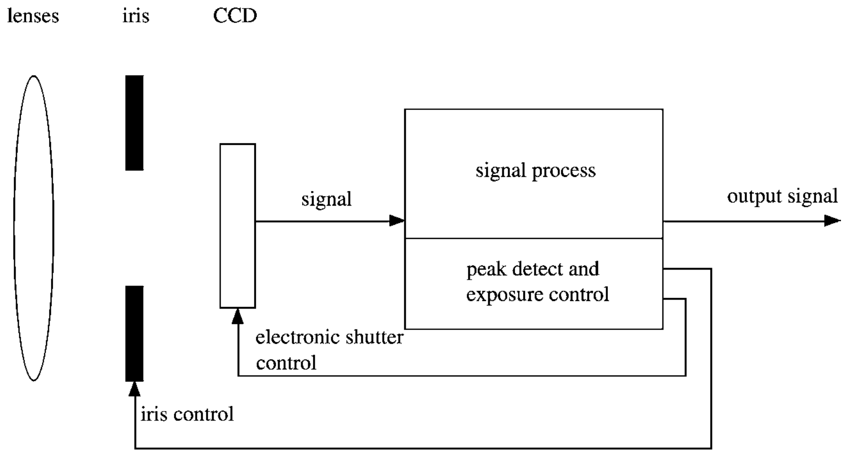


Figure C2.1.73. Automatic exposure.

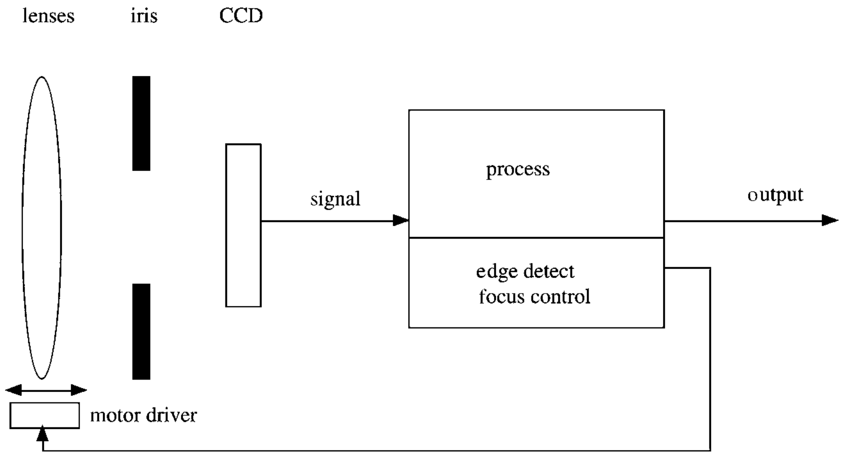

lenses
iris
CCD
signal
process
edge detect
focus control
output
motor driver

Figure C2.1.74. Automatic focus.

Automatic image stabilization

This function reduces image blurring during, for example, hand-held camera shooting. Detection may be performed either electronically (from the image) or mechanically, and correction likewise may be performed either electronically (by image processing) or mechanically [46]. Combinations of these detection and correction methods can be used to configure a system. In the example shown in figure C2.1.75, camera shaking is detected mechanically while correction is performed electronically.

Electronic zoom

The electronic zoom function expands a portion of the picture electronically much like extending the range of a zoom lens [47]. This function is achieved by storing each frame in a memory, controlling

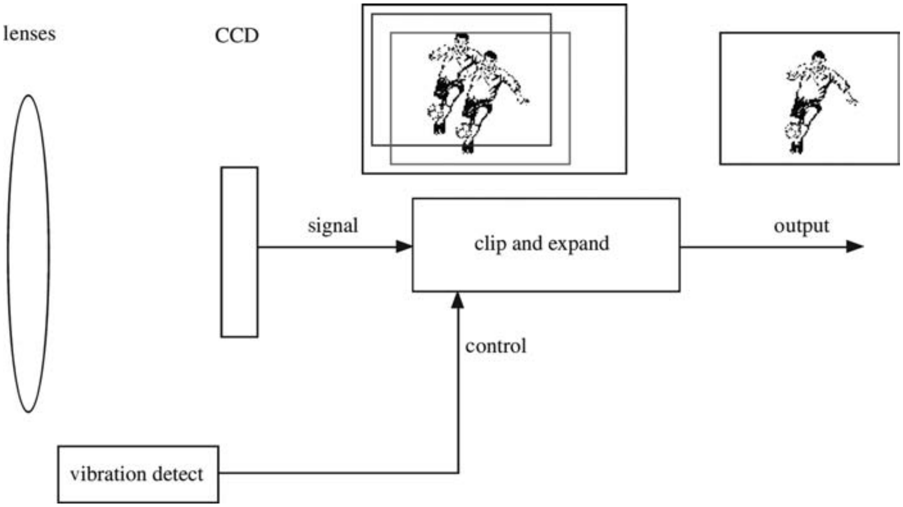

lenses
CCD
signal
clip and expand
output
control
vibration detect

Figure C2.1.75. Image stabilizer.

the position to be read out and readout speed, and performing interpolation with regard to pixels that are lacking.

Defect compensation

This function compensates for CCD pixel defects. This is done by storing the defect position in ROM beforehand and then replacing a defective pixel by a neighbouring pixel in accordance with drive timing.

References

[1] Weimer P K, Forgue S V and Goodrich R R 1950 The Vidicon photoconductive camera tube *Electronics* **23** 71–73

[2] De Hann F F, Van der Drift A and Schampers P P M 1963/64 The 'Plumbicon' a new television camera tube *Philips Tech. Rev.* **25** 133–151

[3] Shimizu K, Yoshida O, Aihara S and Kiuchi Y 1971 Characteristics of experimental CdSe Vidicons *IEEE Trans.* **ED-18** 1058–1062

[4] Shidara K, Goto N, Maruyama E, Hirai T and Nonaka N 1981 The advanced composition of SATICON photoconductive target *IEEE EDL* **EDL-2** 101–102

[5] Maruyama E 1982 Amorphous build-in-field effect photoreceptors *Japan. J. Appl. Phys.* **21** 213–223

[6] Tanioka K, Yamazaki J, Shidara K, Taketoshi K, Kawamura T, Ishioka S and Takasaki Y 1987 An avalanche-mode amorphous selenium photoconductive layer for use as a camera tube target *IEEE* **EDL-8** 392–394

[7] Robinson G A 1977 The silicon intensifier target tube *SMPTE J.* **86** 414–418

[8] Pai D M and Enck R C 1975 Onsager mechanism of photogeneration in amorphous selenium *Phys. Rev. B* **11** 5163–5174

[9] Okano F, Kumada J and Tanioka K 1990 The HARP high-sensitivity handheld HDTV camera *SMPTE J.* **99** 8

[10] Kubota M *et al* 1996 Ultrahigh-sensitivity new super-HARP camera *IEEE Trans. Broadcasting* **42** 251–258

[11] Tanioka K, Ohkawa Y, Miyakawa K, Suzuki S, Takahata T, Egami N, Ogusu K, Kobayashi A, Hirai T and Kawai T 2001 Ultra-high-sensitivity new super-HARP pickup tube *IEEE Workshop on CCD and Advanced Image Sensors, June 7–9* pp 216–219

[12] Neuhauser Robert G 1987 Photoconductors utilized in TV camera tubes *SMPTE J.* **96** 473–484

[13] Horton J W *et al* 1964 The scanister—a solid-state image scanner *Proc. IEEE* **52** 1513–1528

[14] Weimer P K *et al* 1964 A self-scanned solid-state image sensor *Proc. IEEE* **55** 1591–1602

[15] Boyle W S and Smith G E 1970 Charge coupled semiconductor devices *Bell Systems Tech. J.* **49** 587–593

[16] Mendis S M, Kemeny S E and Fossum E R 1994 CMOS active pixel image sensor *IEEE Trans. Electron Devices* **41** 452–453

[17] Weckler G P 1967 Operation of pn junction photodetectors in a photon flux integration mode *IEEE J. Solid-State Circuits* **SC-2** 65–73

[18] Kosonocky W F and Carnes J E 1973 Two phase charge coupled devices with overlapping polysilicon and aluminum gates *RCA Rev.* **34** 164–202

[19] Ishihara Y *et al* 1983 A high photosensitivity IL-CCD image sensor with monolithic resin lens array *International Electron Device Meeting Technical Digest* pp 497–500

[20] Yamada T *et al* 2000 A progressive scan CCD image sensor for DSC applications *IEEE J. Solid State Circuits* **SC-35** 2044–2054

[21] Smith C *et al* 1999 An 8M-CCD for an ultra high definition TV camera *1999 IEEE Workshop of Charge Coupled Devices and Advanced Image Sensor* pp 175–178

[22] Nobusada *et al* 1989 Frame interline transfer CCD sensor for HDTV camera *International Solid-State Circuit Conference, Dig. Tech. Papers* pp 88–89

[23] ItakuraK *et al* 1997 A 2/3-in 2.0M-pixel CCD imager with an advanced MIFT architecture capable of progressive scan *IEEE Trans. Electron Devices* **44** 1625

[24] Yonemoto K and Sumi H *et al* 2000 A CMOS image sensor with a simple fixed-pattern-noise-reduction technology and a hole accumulation diode *IEEE J. Solid State Circuits* **35** 2038–2043

[25] Inoue S *et al* 2001 A 3.25M-pixel APS-C size CMOS image sensor *2001 IEEE Workshop of Charge Coupled Devices and Advanced Image Sensors* p 16

[26] Loose M *et al* 2001 2/3 CMOS imaging sensor for high definition television *2001 IEEE Workshop of Charge Coupled Devices and Advanced Image Sensors* p 44

[27] Krymski A *et al* 2001 A high speed, 240 frame/s, 4megapixel CMOS image sensor *2001 IEEE Workshop of Charge Coupled Devices and Advanced Image Sensors* p 28

[28] Burkey BC *et al* 1984 The pinned photodiode for an interline-transfer CCD image sensor *International Electron Device Meeting Technical Digest* pp 28–31

[29] Carnes J E and Kosonocky W F 1972 Noise source in charge coupled devices *RCA Rev.* **33** 327–343

[30] White M H *et al* 1974 Characterization of surface channel CCD image arrays at low light levels *IEEE J. Solid-State Circuits* **SC-9** 1–13

[31] Sequin C H 1972 Blooming suppression in charge coupled area imaging devices *Bell System Tech. J.* **51** 1923–1928

[32] Ishihara Y *et al* 1982 Interline CCD image sensor with anti-blooming structure *International Solid-State Circuits Conference Technical Papers* pp 168–169

[33] Toyoda A *et al* 1991 A novel tungsten light-shield structure for high-density CCD image sensors *IEEE Trans. Electron Devices* **ED-38** 965–968

[34] Teranishi N *et al* 1982 No image lag photodiode structure in the interline CCD image sensor *International Electron Device Meeting Technical Digest* pp 324–327

[35] Seib D H 1974 Carrier diffusion degradation of modulation transfer function in charge coupled imagers *IEEE Trans. Electron Devices* **ED-21** 210–217

[36] Vanstone G F, Roberts J B G and Long A E 1974 The measurement of the charge residual for CCD transfer using impulse and frequency responses *Solid-State Electron.* **17** 889–895

[37] Hiroshima Y *et al* 1984 Elimination of fixed pattern noise in super-8 format CCD image sensor by use of epitaxial wafers *International Electron Device Meeting Technical Digest* pp 32–35

[38] Jastrzebski L *et al* 1984 Silicon wafers for CCD imagers *J. Electrochem. Soc.* **134** 212–221

[39] Lang H and Bouwhuis G 1962 Colour separation in colour television cameras *Philips Tech. Rev.* **24** 263

[40] Bayer B E 1976 Color imaging array *U.S. Patent* 3,971,065.

[41] Kuriyama T, Kodama H, Kozono T, Kitahama Y, Morita Y and Hiroshima Y 1991 A 1/3-in 270000 pixel CCD image sensor *IEEE Trans. Electron Devices* **38** 949–953

[42] White M H, Lampe D R, Blaha F C and Mack I A 1974 Characterization of surface channel CCD image arrays at low light levels *IEEE J. Solid-State Circuits* **SC9** 1–13

[43] Recommendation ITU-R BT.470-6 1998 *Conventional Television Systems*

[44] Recommendation ITU-R BT.601-5 1995 *Studio Encoding Parameters of Digital Television for Standard 4:3 and Wide-Screen 16:9 Aspect Ratios*

[45] Recommendation ITU-R BT.709-4 2000 *Parameter Values for the HDTV Standards for Production and International Programme Exchange*

[46] Oshima M, Hayashi T, Fujioka S, Inaji T, Mitani H, Kajino J, Ikeda K and Komoda K 1989 VHS camcorder with electronic-image stabilizer *IEEE Trans. Consumer Electron.* **35** 749–758

[47] Pulford T P, Risk R J, Balmer J and Broadberry R 1980 An electronic zoom facility for television *Electron. Eng.* **52** 20

Further reading

Canon, Inc. 1992 *TV Optics II The Canon Guide Book of Optics for Television System* (Kawasaki: Canon Inc.)

Luther A C 1998 *Video Camera Technology* (Norwood, MA: Artech House)

Ochi S *et al* 1996 *Charge-coupled Device Technology* (London: Gordon and Breach)

Whitaker J and Benson B 2000 *Standard Handbook of Video and Television Engineering* 3rd edn (New York: McGraw-Hill)

C2.2
Vacuum tube and plasma displays

Makoto Maeda, Tsutae Shinoda and Heiju Uchiike

C2.2.1 Vacuum tube devices

C2.2.1.1 CRT structure and its operation

The cathode ray tube (CRT) is a luminescent display invented by K F Braun in 1897. The display, which is inexpensive but can show resolute pictures on its screen, has been the leading technology in the display field over more than 100 years. But the heavy weight and long depth are the CRT's weaknesses. The liquid crystal display (LCD) and other new flat panel displays that have no weaknesses as such are expanding their market share in recent years.

The CRT comes in several types—direct view, monochrome and projection. Every year, as many as 200 million pieces are produced and the demand is steadily growing by 3%. Based on the television broadcast specifications, the CRT with the screen's length–height ratio of 4:3 was the most common but the CRT with the ratio of 16:9 is now popularized. Even the 1:1 ratio CRT is produced for some special customers. The CRT screen was once round to intensify the glass strength but a series of recent technological innovations has made it possible to design more varieties of flat-screen-type CRT. To continue to be the leader in display business, the CRT needs to be more cost-competitive, solve the weight and depth problems and even enhance its strengths—brightness, colour and contrast—to produce more beautiful pictures.

CRT operation

Figure C2.2.1 shows the monochrome CRT structure. Video signal is fed to the cathode which is a part of the electron gun. The cathode generates free electron. The electron is focused by the electron gun like light is focused by a lens. Electrons travel freely in its evacuated inside, while glass is used to 'envelope' the CRT. Electro-conductive substance is applied to the funnel's inside to form a film that keeps the inside's electric potential stable. The film builds as a high-voltage condenser in-between with another film made on the outside of the funnel. The condenser stabilizes the supply of anode voltage. Emitted by the electron gun and accelerated with the voltage of 20–30 kV, electrons travel fast in the form of a beam onto the panel coated with phosphor. The deflection yoke creates a magnetic field that bends the electron beams and makes them scan the entire panel. The electrons hit against the phosphor that emits light. Individual components and their function are described below.

Electron gun

Figure C2.2.2 shows the cross-section of the electron gun. This device is equipped with a cathode and operates in an electric lens system. The cathode is usually an oxide composition of barium (Ba),

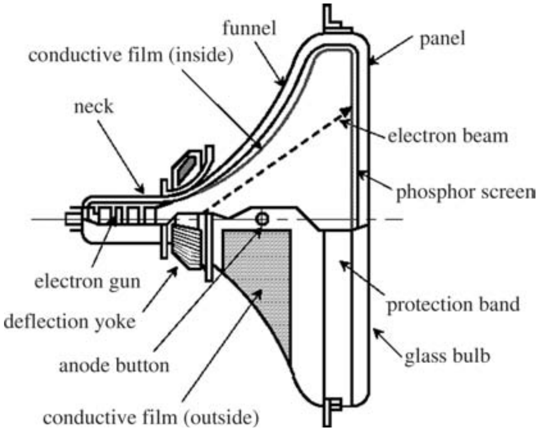

funnel
conductive film (inside)
panel
neck
electron beam
phosphor screen
electron gun
deflection yoke
protection band
anode button
glass bulb
conductive film (outside)

Figure C2.2.1. CRT structure.

calcium (Ca) and strontium (Sr). Activated by heat at about 800°C, the cathode emits electrons. When varying picture signal voltage with an amplitude of about 100 V is applied to the cathode, the volume of electrons going through the grit 1 (G 1) changes: the bigger the volume, the brighter the picture and vice versa. But to focus the dispersed electrons emitted from the cathode in large current and to create pictures on the panel, we must carefully design the structure and shape of the gun's electrode and the arrangement of applied voltage. Using 100 V signals, the electron gun can control electrons of 30 keV of energy. This function as a noise free amplifier is unique and not found in any other device of flat panel displays.

Deflection

An electron beam put out from the gun travels straight toward the centre of the CRT. To let it reach all over the screen, we usually use two methods—electrostatic deflection and electromagnetic deflection. The former method requires two flat-plate electrodes facing each other. The beam travels between the two. When electrical potential between them varies, the beam changes the direction. In spite of its lower

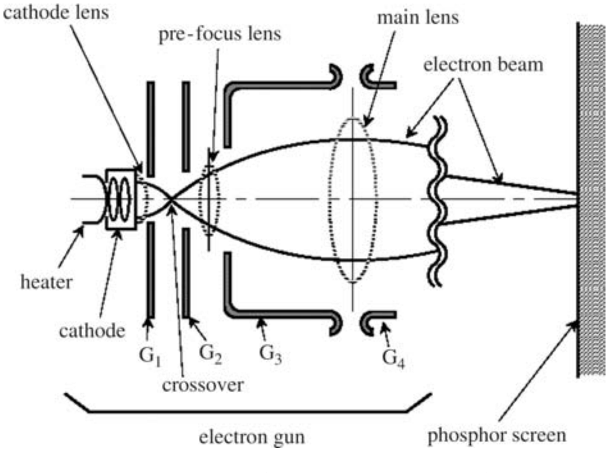

cathode lens
pre-focus lens
main lens
electron beam
heater
cathode
G_1
G_2
G_3
G_4
crossover
electron gun
phosphor screen

Figure C2.2.2. Electron gun cross-section.

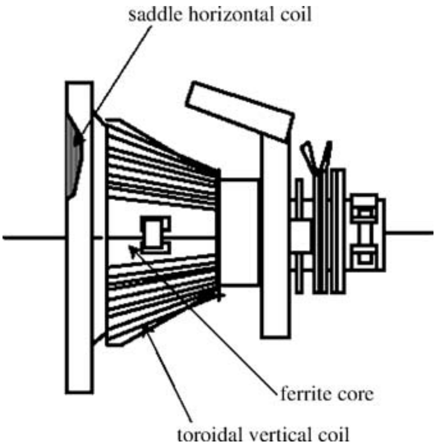


Figure C2.2.3. Deflection yoke.

deflection efficiency, the method is quite effective in deflecting high frequency. Meanwhile, the latter method of electromagnetic deflection characterized by its higher deflection efficiency is used for television and many other CRTs, and the deflection yoke performs electromagnetic deflection, using two pair of coils. As a pair of coils facing each other generates one magnetic field, the two pairs create two magnetic fields, directing the beam both horizontally and vertically. See the structure in figure C2.2.3 for magnetic deflection yoke.

Phosphor screen

The inside of the CRT panel is coated with layers of phosphor particles. Each particle is 3–10 μm in diameter. The aluminium film covers and protects the phosphor. Gasses still remaining in the CRT would be hit by electrons and become ionized. Without the film, the ion would crash against the phosphor screen and damage the phosphor. This film is also effective in raising CRT luminance by reflecting light coming from the phosphor. It stabilizes electric potential around the screen as well.

C2.2.1.2 Monochrome CRT

Monochrome rather than colour CRTs are mainly used in the medical field where high resolution and high brightness pictures are required. The electromagnetic focusing method is applied to the electron gun to achieve high resolution. Electric current running through the coils attached around the CRT neck induces magnetic field and works as electro-magnetic lens. Being put outside the neck, the 'lens' has a very small aberration, making the beam spot extremely tiny.

C2.2.1.3 Projection CRT

The colour picture projection CRT is a combination of three monochrome CRTs in red, green and blue, respectively. Pictures on each of the three single CRT panels are expanded through each optical lens and projected onto the outside screen where the pictures are combined. The system is depicted in figure C2.2.4. The phosphor screen of the single CRT is usually 7–9 in. in diagonal. The size of projected pictures is as large as 40–60 in. in diagonal, and higher brightness (10 000 cd m^{-2}) and higher resolution (the spot size is 0.2 mm in diameter) are requisite. If larger electric current is applied to illuminate phosphor, however, the phosphor becomes extremely hot; CRT brightness is saturated; and phosphor

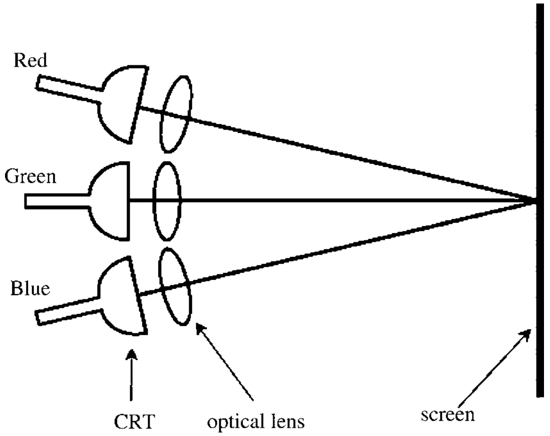


Figure C2.2.4. CRT projection system.

quality degrades. Then, the CRT fails to operate. To prevent this problem, the CRT panel is kept cool with coolant, while a kind of phosphor strong enough to bear such heavy load is applied.

C2.2.1.4 Colour CRT

Principle

Phosphor in three different colours is applied to produce colour pictures on the CRT screen. Figure C2.2.5 shows the mechanism for exciting phosphor. The colour selection device (mask) is put in front of the phosphor-coated panel. Three pieces of cathode in the electric gun put out electrons, which travel through the mask and hit onto red, green and blue phosphor and let it illuminate.

Colour selection mechanism

Among several types of colour selection masks, the shadow mask and the aperture grille are widely applied to the CRT these days. The former is steel plate of 0.2 mm in thickness with round- or

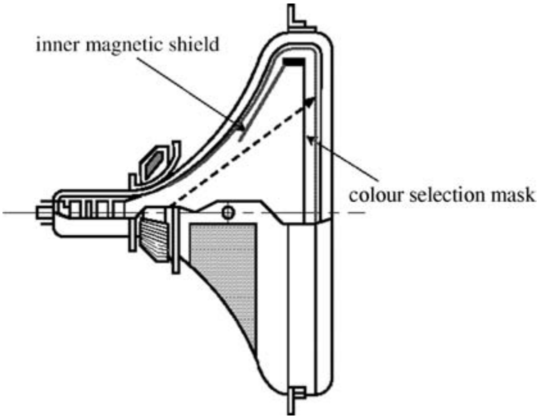


Figure C2.2.5. Colour CRT structure.

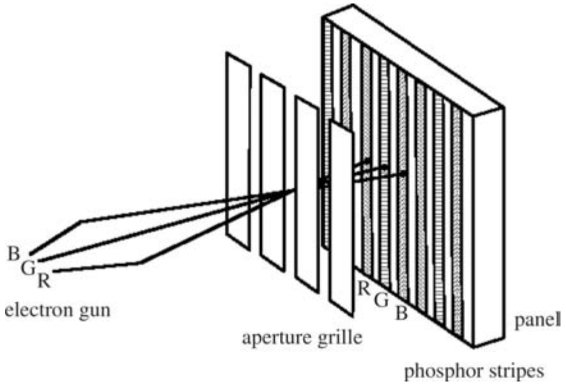


Figure C2.2.6. Aperture grille system.

rectangular-shaped holes 0.2–1.0 mm apart. The mask's opening must be put in place to let a beam from one particular colour cathode reach the same colour phosphor. Hit by electron beams, the mask becomes hot and expands and the position of the opening moves. Therefore, the material Invar that has the very small thermal expansion coefficient is sometimes used for the mask. To match more varieties of flat-faced CRT designed these days, new shadow masks are being developed.

The aperture grille has a shape of a vertical reed screen. Both the shadow mask and the aperture grille have light transmittance of about 20%. This means 80% of the electrons out of the gun do not go through the mask. Figure C2.2.6 shows the aperture grille system and figure C2.2.7 shows the shadow mask system.

Phosphor screen

Slurry that contains phosphor and photo sensitizer is applied to the panel to create the phosphor screen. After being dried, the coated slurry is exposed to ultraviolet (UV) radiation light emitted from the lamp. When the panel is washed with water, the material except the light-exposed and hardened phosphor flows away. After this process is repeated three times, red, green and blue phosphor stripes appear; black stripes between each different colour phosphor are also created by a similar method. Figure C2.2.8 illustrates the cross-section of the phosphor screen.

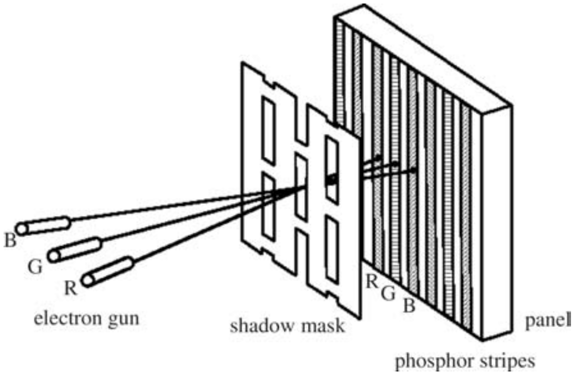


Figure C2.2.7. Shadow mask system.

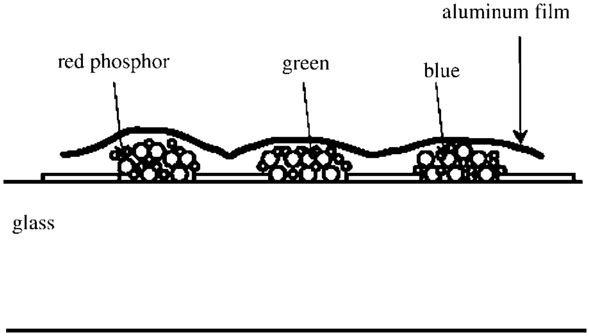


Figure C2.2.8. Phosphor screen cross-section.

Electron gun

The modern electron gun for a colour picture tube is quite complicated compared to the basic structure shown in figure C2.2.2. Figure C2.2.9 depicts an example of the electron gun for colour CRT [6]. As one gun is equipped with one piece of cathode and its electrons are designated to illuminate the corresponding colour phosphor, the colour CRT generally requires three electron guns to illuminate three colours to produce colour pictures.

The trinitron gun is equipped with one large electric lens for three electron beams. A large lens generally shows better performance than a small lens.

Deflection yoke

Unlike the monochrome CRT, the colour CRT needs to deflect three electron beams at once. Those beams are required to converge into every spot spread all over the phosphor screen in order to realize quality pictures. The deflection yoke helps distribute needed magnetic field.

Purity

The CRT is designed for electron beams to travel through the mask and hit the designated phosphor. But this is disturbed when the path is affected by terrestrial magnetism. To prevent this problem, the

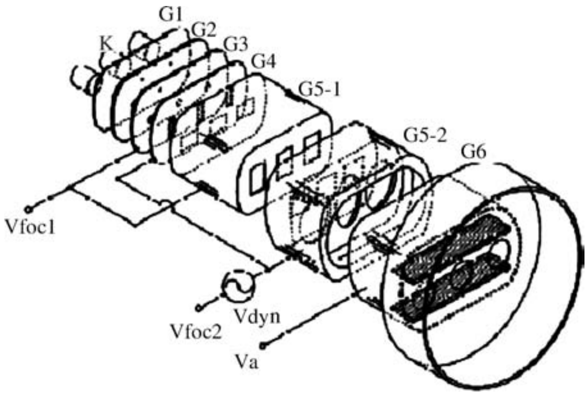


Figure C2.2.9. Electron gun for colour CRT by Wada *et al* [6].

CRT is equipped with a magnetic shield. When the CRT power switch is turned on, attenuating alternating current runs through the coil installed around the CRT to produce attenuating magnetic field. This 'degauss' process is designed to magnetize the shield in the intensity opposite to that of the outside magnetic field and to alleviate the influence.

C2.2.1.5 Contrast

Several measures are taken to prevent ambient light reflection off the surface of the panel and the phosphor screen and to maintain good picture contrast. For instance, the swath between colour stripes (dots) is blackened to halve the light reflection without blocking phosphor light.

Glass of lower transmittance is applied to the CRT panel. Incident rays come through the panel glass, reflect against the phosphor screen and go out through the glass again. Glass of low light transmittance can reduce light reflected through the two passages substantially. Phosphor light travels through the glass but the passage is only once. As the lower the glass transmittance, the better the picture contrast, the transmittance rate of about 50% is the most favourable (see figure C2.2.10).

However, 4% of incoming light is still reflected at the surface of the panel. Computer display CRTs are equipped with glass whose surface is treated in the non-reflection process.

C2.2.1.6 Safety

Glass is used for the CRT and air in its inside is exhausted. Being exposed to the atmospheric pressure, even a small fault of the CRT could lead to a dangerous implosion, dispersing glass pieces all around. A metal band is applied around the CRT to prevent such accidents. The band that is a little smaller in circumference than the CRT reinforces the glass strength by cancelling out glass stress caused by the atmospheric pressure. The band prevents damage even if the glass is broken.

C2.2.1.7 Other CRTs

Flat CRT

A big challenge for the CRT is how to shorten the long depth. As the deflection yoke bends electron beams, the CRT depth becomes shorter if the yoke's deflection angle is wider. Most TV CRTs these days have the yoke with the angle of 110° and new CRT models even wider 120°. But the larger deflection angle requires much more deflection power to bend the beam more sharply. The magnetic field gets distorted and the beams running through it end up producing poor pictures.

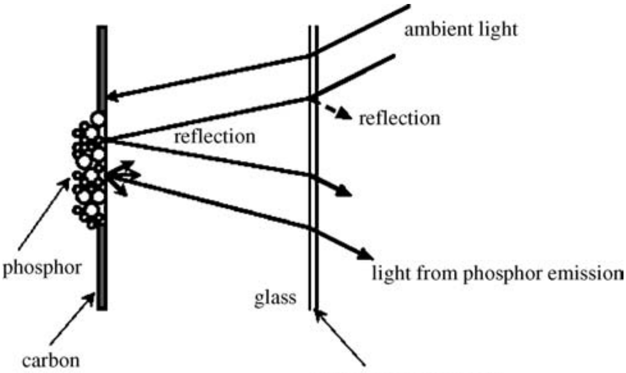


Figure C2.2.10. Light from phosphor emission and ambient light reflection.

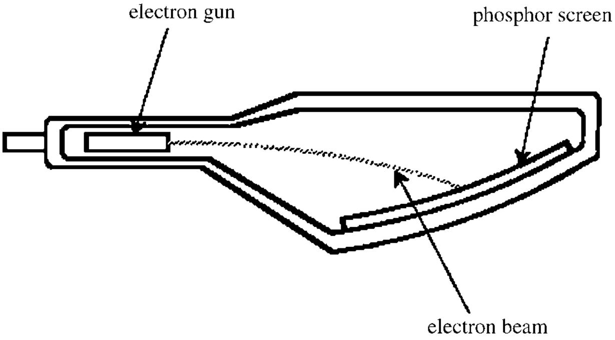


Figure C2.2.11. Flat monochrome CRT.

Maeda [7] designed a thin monochrome CRT (figure C2.2.11). So far, no colour CRT adopting this technology has been sent to the market.

Unlike the traditional CRT, the field emission display (FED) is shorter in depth. It has no deflection device. Now, engineers are intensifying R&D effort to produce this type of product too.

Beam index CRT

Only 20% of electrons are transmitted through the colour selection mask and the rest are absorbed into the mask. Many yeas ago, the colour CRT without the mask to raise power efficiency was developed and sold in the market. This CRT is designed to make one electron beam scan all over the phosphor screen. The beam should have a very small beam spot size not to strike more than one stripe of phosphor. That stripe stands next to another stripe of phosphor, which emits none of the three colours but UV as a signal when the beam moves onto it. Detecting that signal and finding its own location, the beam immediately changes its picture colour information signal. The problem of this system is that the beam spot size becomes bigger when high beam current for high luminance is generated. CRT engineers want to solve this problem and succeed in developing this CRT some day. Their effort will continue until that day (see figure C2.2.12).

C2.2.1.8 Recent developments

Efforts to improve performance and cost reduction are continuing. Some of the recent developments are described below.

Okano developed a 21-in. CRT having very high resolution. Aperture grille pitch at the screen centre of CRT measures 0.126 mm. Horizontal resolution of 2800 dots was realized [8].

Beam index tube is a CRT which has no mask. A new idea was proposed by Bergman *et al* [9]. Primary function of the mask, colour selection, is taken over by an electronic control system. This CRT (called F!T tube) employs the system which has phosphor stripes parallel to electron beam scanning lines. Figure C2.2.13 shows the tracking principle of this new beam index tube.

One of the major problems of CRT is its weight. Most of the weight comes from the glass envelope. Sugawara [10] reduced funnel weight by redesigning the shape of the funnel shown in figure C2.2.14.

The cathode is another component that needs to improve its performance. Oxide cathode is used for most CRTs. The current density from oxide cathode is limited. The new cathode called a hopping electron cathode (figure C2.2.15) was proposed by van der Varrt *et al* [11]. It is based on a self-regulation secondary emission process enabling transport of electrons over insulation surface. The cathode utilizes this mechanism to compress electrons coming from a large conventional cathode into a small funnel

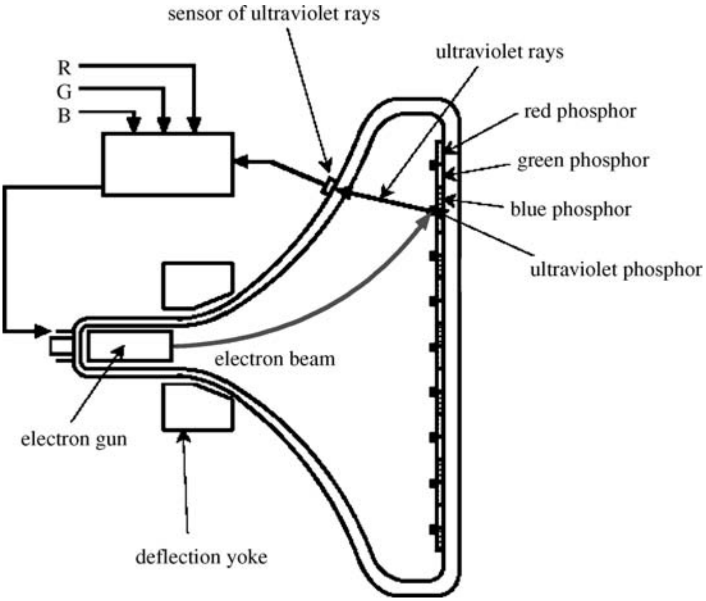

sensor of ultraviolet rays
ultraviolet rays
R
G
B
red phosphor
green phosphor
blue phosphor
ultraviolet phosphor
electron beam
electron gun
deflection yoke

Figure C2.2.12. Beam index CRT.

structure of insulating material. The exit of the funnel serves as a high-brightness electron source for a CRT and can be used to reduce the spot size.

C2.2.2 Plasma display

C2.2.2.1 Introduction

Plasma displays have greatly advanced in the 1990s and are getting a position in the mainstream of the large area flat panel television and display. The road to the development was long and not peaceful

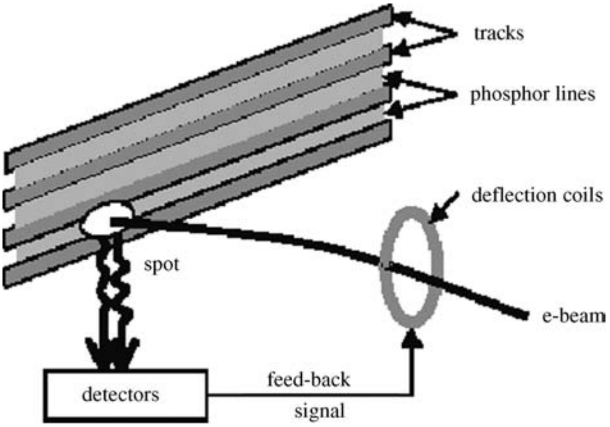

tracks
phosphor lines
deflection coils
spot
e-beam
detectors
feed-back
signal

Figure C2.2.13. Tracking principle of F!T tube by Bergman *et al* [9].

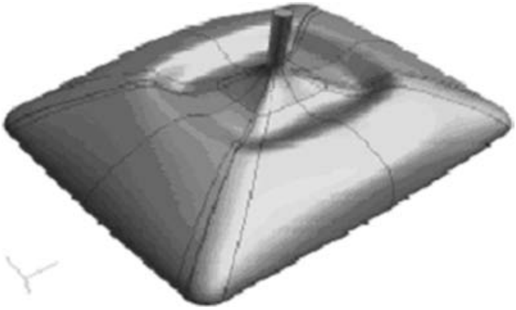


Figure C2.2.14. Novel funnel shape to reduce its weight by Sugawara *et al* [10].

in 30 years. Both successes of the colour moving picture presentation on plasma display panels [12, 13] and production of a 42-in. diagonal PDP [14] promise great business growth. The further developments of the interlaced and progressive displays from 32 to 60 in. HDTVs are making a new value-added market in addition to the market replacing the conventional TV of CRT because PDP has made a new large area and beautiful displays possible and is giving people a new impression.

C2.2.2.2 Development of colour plasma displays

There were two ways researched to present colour image on PDPs. One was to use visible light caused by the discharge. The other was to use visible light from excited phosphors by UV rays or electrons in discharge. As a result of these researches, phosphor system with UV ray excitation has been applied to the recent colour PDPs, because the system was superior to the other methods due to high luminance and colour purification. In particular, the gas system of Ne and Xe including 4–5% of Xe contents showed the excellent results to achieve high luminance and luminous efficiency. Vacuum UV rays of 147, 152 and 172 nm are radiated from Xe and Xe-dimers as shown in figure C2.2.16.

At the early stage of colour plasma display development, both AC and DC PDPs were carried out to accomplish computer monitors and colour television. Most of the colour PDP researches to achieve colour television were DC ones, because colour AC PDPs had difficulty in attaining long lifetime due to the degradation of colour phosphors caused by the ion sputtering [15]. Because phosphors in DC colour PDPs were deposited around the anodes and then no ion bombardment, colour DC PDPs had an advantage of longer phosphor lifetime compared to that of AC ones at that stage.

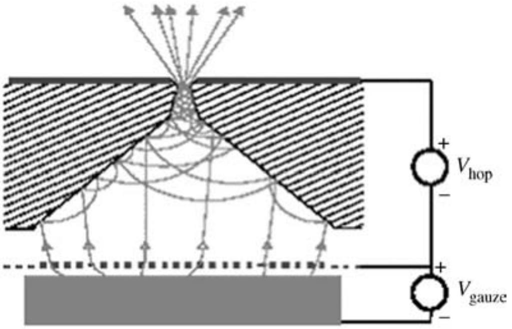


Figure C2.2.15. Hopping electron cathode by van der Vaart *et al* [11].

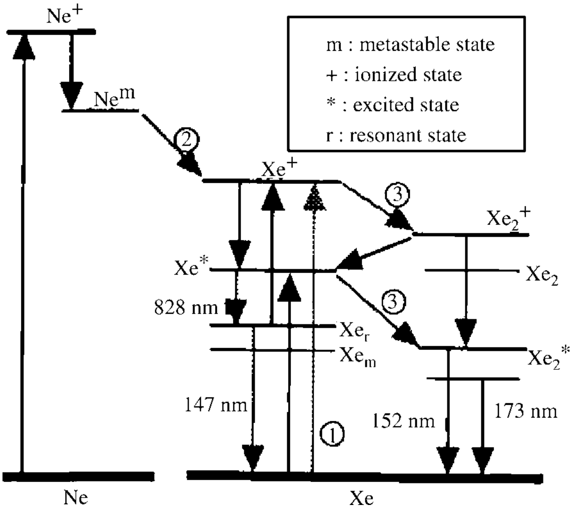


Figure C2.2.16. Energy level of transition for Ne + Xe mixture gas.

Colour DC PDPs for television were investigated by the application of negative glow and positive column region. Even though the colour DC PDPs using positive column had higher luminous efficiency compared to those of negative glow region, the luminance and luminous efficiency were still not sufficient to be released into the market.

In 1983, NHK developed 16-in. diagonal colour DC PDPs whose luminance and luminous efficacy were 21 cd m^{-2} and 0.05 lm W^{-1}, respectively [16]. The performance of luminance and luminous efficacy by NHK was not able to cross the value of 150 cd m^{-2} and 0.4 lm W^{-1}, respectively [17].

The practical panel structure for a colour plasma display panel is called a three-electrode surface discharge belonging to a kind of AC-type plasma display [18]. AC PDP was invented by Bitzer and Slottow in 1966 [19] and the monochrome display has been put into practical use with opposed discharge technologies. The colour PDP, however, did not succeed in practical use although there were many researches with the opposed discharge technologies. There were evolutional developments in structural and operational technologies from monochrome PDP to put the colour AC PDP into practical use. The new direction to develop the colour PDP was opened by introducing surface discharge technologies.

Table C2.2.1 shows the summary of the development of the colour PDP technologies with surface discharge. Takashima [20] firstly reported the application of surface discharge technology on the colour PDP for segment-type display in 1973. Dick [21] has also reported the surface discharge technology of a matrix-type monochrome PDP in 1974. The matrix-type colour PDP with surface discharge technologies has been proposed and followed by authors from 1979 [22, 23]. From the beginning of the research and development, the surface discharge colour AC PDPs showed extremely excellent luminous efficacy performance of 0.75, 0.4 and 0.15 lm W^{-1} for green, red and blue phosphors, respectively [24]. Figure C2.2.17 shows the comparison between opposed discharge and surface discharge PDPs. The phosphors were deposited on the surface of the MgO layers on the opposed electrodes as shown in figure C2.2.17(*a*) and alternated pulses were applied between the electrodes to ignite the discharge. Then the phosphors exposed to the discharge resulted in the rapid degradation due to the ion bombardment in the discharge. On the other hand, the phosphors were deposited on the front cover glass substrate placed away from the discharge area as shown in figure C2.2.17(*b*); therefore, the surface discharge colour PDPs have got an advantage of long life. All of these surface discharge colour

Table C2.2.1. Development of colour PDP technologies with surface discharges.

Color technologies history

Year	Researcher	Phosphors	Electrode	Electrode on	Type	Electrode configuration	Ribs
1973	Takashima	Green	Two	Single substrate	Transmitting	Segment	Glass sheet
1974	Dick	Non	Two	Single	–	Matrix	Stripe
1979	Shinoda	RGB	Two	Single	Transmitting	Matrix	Non
1984	Shinoda	RGB	Three	Single	Transmitting	Matrix	Stripe + mesh
1985	Dick	Non	Three	Double	–	Matrix	
1987	Shinoda	RGB	Three	Double	Reflecting	Matrix	Stripe + mesh
1992	Shinoda	RGB	Three	Double	Reflecting	Matrix	Stripe

PDPs employed two electrodes. The two-electrode surface discharge system, however, did not succeed in developing the practically available colour PDPs.

The research has finally resulted in a new structure with the three electrodes as shown in Table C2.2.1 [17, 25]. Shinoda invented both essential technologies such as the three-electrode structure and a new greyscale driving technology, or address-, display-period separation method (ADS method), which enabled the realization of practical colour plasma displays. Figure C2.2.18 summarizes the technical issues to develop the colour PDP, such as realizations of colour PDP, long operating life, high luminance, high resolution and full colour operation when the research for the surface discharge colour PDP was started in 1979. Developing a three-electrode panel structure has solved these first four issues. And development of the new driving technology has solved the last one. Finally, the three-electrode PDP structure with stripe rib and phosphor structure has been completed and the practical 21-in. diagonal colour PDP has been developed with these technologies in 1992 as shown in figure C2.2.19. The larger 42-in. diagonal plasma display shown in figure C2.2.20 was put onto the market in 1996, which started the era of plasma television.

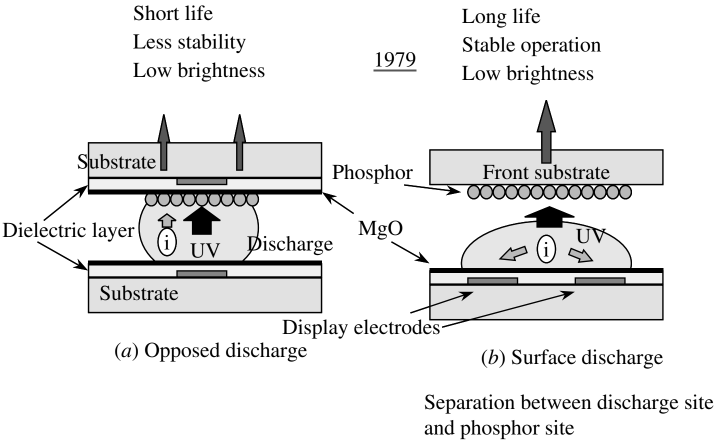

Short life
Less stability
Low brightness
1979
Long life
Stable operation
Low brightness
Substrate
Phosphor
Front substrate
Dielectric layer
i
UV
Discharge
MgO
UV
i
Substrate
Display electrodes
(a) Opposed discharge
(b) Surface discharge
Separation between discharge site
and phosphor site

Figure C2.2.17. Comparison between surface discharge and opposed discharge. Phosphors are deposited on the front cover glass substrate placed away from the discharge area in the surface discharge.

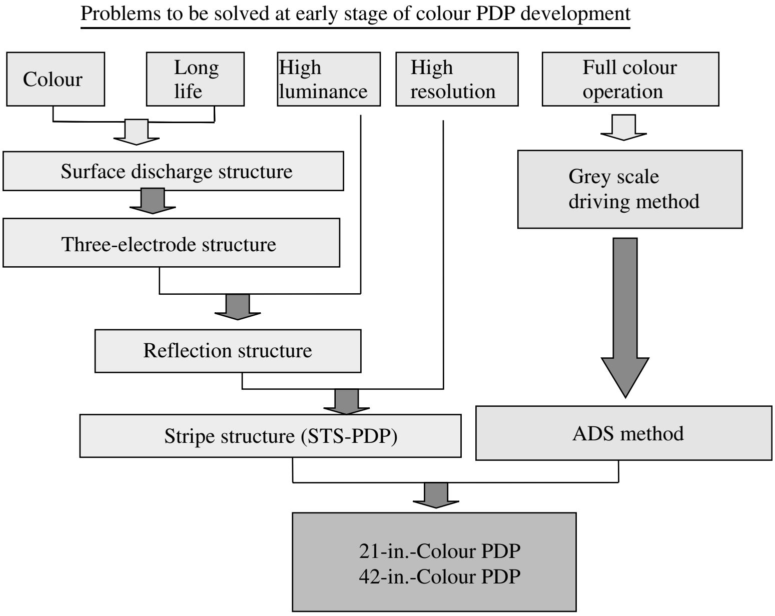

Problems to be solved at early stage of colour PDP development
Colour
Long life
High luminance
High resolution
Full colour operation
Surface discharge structure
Three-electrode structure
Grey scale driving method
Reflection structure
Stripe structure (STS-PDP)
ADS method
21-in.-Colour PDP
42-in.-Colour PDP

Figure C2.2.18. Summary of the technical issues to develop the surface discharge colour PDP.

C2.2.2.3 Essentials of colour AC PDP

Operating principle of colour PDP

The colour PDP is the display using the luminance from the combination of the phosphors and gas discharge. It is possible to think of a colour PDP model as if some millions of miniature florescent lamps are arranged between the glass plates of an area of 1 m^2. Figure C2.2.21 shows the luminance model in the each lamp, which is usually called a cell. When the voltage is applied to the gas, discharge is ignited in each cell making ions and electrons from the atoms. The ions and electrons lose energy by emitting a

(1992)

Item	Performance
Display Area	422 × 316 mm
Aspect Ratio	4:3
Number of Pixels	640(R,G,B) × 480
Pixel Pitch	0.66 mm × 0.66 mm
Number of Colours	260,000
Luminance	180 cd m^{-2}
Viewing Angle	>160°
Power Consumption	100 W_{max}
Weight	4.8 kg

Figure C2.2.19. Specification of 21-in. diagonal colour plasma display.

(1996)

Item	Performance
Display Area	920 mm x 518 mm
Aspect Ratio	16:9
Number of Pixels	852 (R,G,B) X 480
Pixel Pitch	1.08 mm X 1.08 mm
Number of Colours	16.7 million
Luminance	$350\ \mathrm{cd\ m}^{-2}$
Viewing Angle	> 160°
Power Consumption	$300\ \mathrm{W}_{\mathrm{max}}$
Weight	18Kg

Figure C2.2.20. A 42-in. diagonal colour plasma display.

UV ray. The plasma display is designed to irradiate UV rays of Xe resonance emission (147 nm) and Xe molecular emission (173 nm). The irradiated UV rays stimulate the phosphors and visible lights are emitted. The three prime colour phosphors are arranged in each of the discharge cells that are at the cross point of the electrodes.

To display a colour image on PDP, the discharge needs to be controlled between the ON and OFF states in each discharge cell (sub-pixel) that has the three prime colour phosphors. The operating principle is simply explained with the opposed discharge structure without phosphors as shown in figure C2.2.22. The electrodes are arranged orthogonally on each opposite glass plate and covered by a dielectric layer. The dielectric on the front substrate is also covered by an MgO protecting layer. The plates are assembled with a gap of about 100 μm and a Ne + Xe gas system is introduced between them.

The operating waveform is composed of the write pulses, the sustain pulses and the erase pulses. The write pulse whose voltage is higher than the firing voltage (V_f), i.e. a threshold voltage to ignite the discharge, is applied to the X electrode and then the discharge is ignited. The ions and electrons generated by the discharge are absorbed to the opposed dielectric surfaces by the electric fields. As a result, the internal electrical field of the cell is reduced rapidly by these absorbed charges and then the

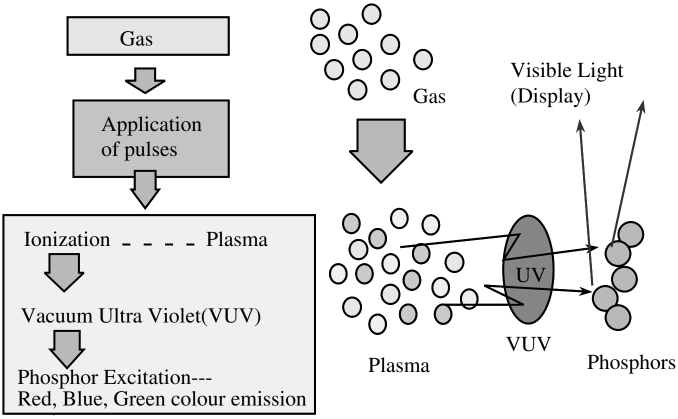
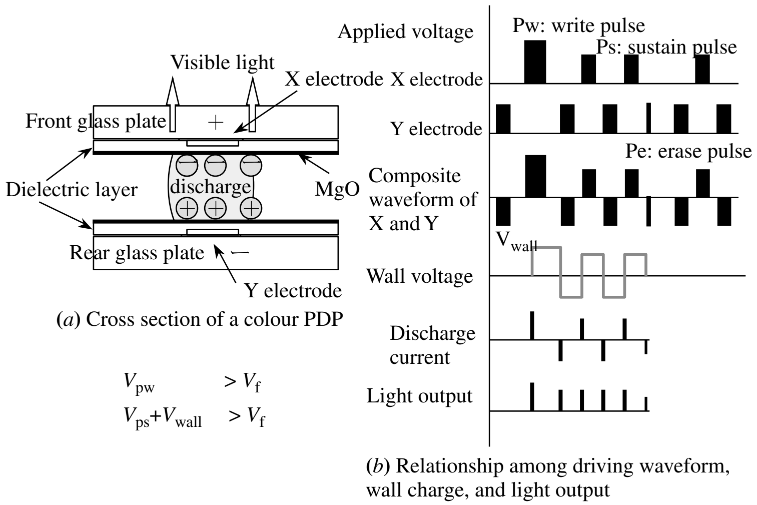


Figure C2.2.21. The luminance model of each cell in plasma display.



(*a*) Cross section of a colour PDP

(*b*) Relationship among driving waveform, wall charge, and light output

Figure C2.2.22. The operating principle with the opposed discharge structure without phosphors.

discharge is stopped. Therefore, the discharge has a pulse shape. The charge deposited on the dielectric is called a wall charge and the voltage through the capacitance between the surface of the dielectric and the electrode under the dielectric is called a dielectric voltage. The voltage due to the wall charge across the gas is called a wall voltage and the voltage across the gas due to both the wall voltage and externally applied voltage is called a cell voltage. The width of the write pulse is adjusted to accumulate a sufficient wall charge on the dielectrics. The sustain pulses are applied on Y electrodes successive to the write pulse. The polarity of the voltage is reversed to the write pulse and the voltage is superimposed on the formerly accumulated wall voltage and then the cell voltage can exceed the firing voltage to ignite discharge, although the applied sustain voltage itself is lower than the firing voltage. The discharge also has a pulse shape similar to the discharge by the write pulse. In the same manner the successive sustain pulses sustain the discharge state and make display. The luminance of the display is proportional to the number of the sustain pulses and from 30 000 to 50 000 pulses per second are usually applied for sustaining. The narrower width than the sustain pulse is applied to erase the discharge. Although the discharge is ignited, the wall charge cannot be accumulated on the dielectric because of the narrow width of the erase pulse. The successive sustain pulses cannot sustain the discharge because the cell voltage is lower than the firing voltage.

In this manner the discharge state (ON state) and the non-discharge state (OFF state) are maintained. This phenomenon is called the memory effect of AC PDP.

Features of colour AC plasma display panel

The developed colour AC PDP has all the advantages of conventional monochrome AC PDP.

The following are the advantages: (1) good nonlinearity, (2) memory function, (3) high addressing speed, (4) wide viewing angle, (5) high luminance for large area and large display capacity, (6) high contrast ratio, (7) high greyscale, (8) full colour, (9) digital display, (10) flat panel, (11) simple structure and (12) large area.

The good nonlinearity means that the relationship between an applied voltage and the luminance has a clear nonlinearity. That is, when the applied voltage is increased, the discharge current is increased rapidly and then bright illumination begins at a certain voltage, and when the applied voltage is decreased, the discharge current decreases rapidly at a certain voltage and then illumination is eliminated. As the luminance levels are quite different between ON and OFF states, the high quality display with high contrast ratio is possible.

The memory effect is the function to maintain the ON and OFF states on the panel and is a great advantage for realizing a large area or a large display capacity. For example, in the case of the CRT without memory effect, the luminance decreases by increasing the display size or the display capacity because the electron beam stimulating the phosphors stays for a short duration at the phosphor surface of a display spot. In contrast, as all of the cells can be illuminated at the same time when the common sustain pulses are applied to the electrodes in the AC PDP, the high luminance level can be kept not depending on the display size or the display capacity.

The high speed addressing is due to the gas device. As the discharge is finished within 1 μs and the wall charge is accumulated within 2 μs after the pulse is applied, the data input is possible within 2 μs. There is a report that the data input is possible within 1.5 μs for one scan line. As a result, the PDP can display a beautiful moving image with 16.8 million colours.

The PDP is expected to play an important role in the future digital society because it is essentially a digital device. The digital ON and OFF states of the discharge cell are easily controlled by the digital signals of the computer and the network.

Principle of three-electrode plasma display panel

Although the colour PDP has been researched since the end of the 1960s when the AC PDP was invented, it was not successful with the two-electrode structure because of the phosphor and MgO degradation and a difficulty in the operation [26]. One of the key breakthroughs that solved the issues was the introduction of the surface discharge and three-electrode structure. Figure C2.2.23 shows the principal electrode configuration and operation. There are three kinds of electrodes, such as two display electrodes (sustain and scan electrodes) and an address electrode. The sustain electrodes are connected to each other with a common electrode and the scan electrodes are independent. And then the one electrode terminal is added to the conventional two-electrode system. There are two kinds of operation method. One is write-in operation, which means the discharges are ignited to make wall charges between the selected address electrodes and a scan electrode for input display data and then sustain the discharge in the cells to be displayed by applying the sustain pulses between the display electrodes. Another one is erase operation, which means discharges are ignited to make the wall charges in all of the cells between display electrodes, scan and sustain electrodes, along a display line at once and then ignite discharges between the address electrodes and a scan electrode to erase the wall charges resulting in elimination of the discharges in the cells not to be displayed.

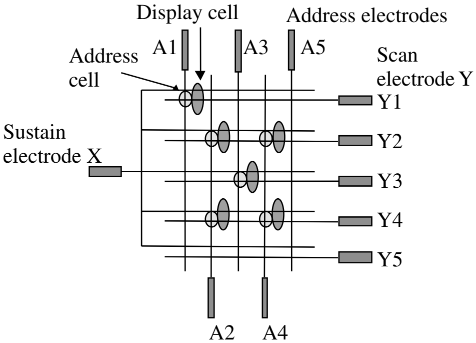


Figure C2.2.23. The principal electrode configuration of three electrode PDP.

C2.2.2.4 Practical panel structure and fabrication process

The first successful full colour PDP has a diagonal size of 21 in. and the fabrication processes were completed while developing the panel. The new simple panel structure with a shape of stripe ribs and phosphors was developed to realize the fine pixel pitch (0.66 × 0.22 mm) for high resolution and was the most suited structure for mass production.

Figure C2.2.24 shows the practical panel structure which is called the three-electrode surface discharge PDP with the stripe structure [27]. Paired parallel display electrodes, sustain electrode and scan electrode, are formed on the front glass substrate. Each display electrode is composed of a transparent ITO (indium–tin oxide) and a narrow bus electrode of multi-layered Cr, Cu and Cr, to emit a luminance effectively through the transparent electrode and reduce the electrode resistance. These electrodes are covered by a dielectric layer, which is made of low melting point glass materials. These are also covered with a thin MgO layer. On the other rear substrate, the striped address electrodes are arranged. Striped barrier ribs are on both sides of the address electrodes to separate the adjacent discharge cells and to eliminate the optical cross talk between them. Three primary colour phosphor materials for red, blue and green colours are deposited in the neighbouring channels made by the ribs to cover both on the side walls of the ribs and on the dielectric layer. The structure has realized good performances such as a high luminance, a high luminous efficacy and a wide viewing angle. Phosphor materials are $BaMgAl_{14}O_{23}$:Eu for blue, $(Y{\cdot}Ga)BO_3$:Eu for red and Zn_2SiO_4:Mn for green. The substrates are assembled to each other with about 150 μm gap. A Ne + Xe gas mixture is introduced between the substrates. The panel structure developed for the 21-in. diagonal colour PDP is the simplest one of conventionally researched colour PDPs. And the fabrication process is also simple enough for mass production. So the PDP has advantages such as a low cost process, and an easiness in the manufacture of the large area panels and the high resolution panels.

The essential fabrication process as shown in figure C2.2.25 is also completed to develop the 21-in. PDP. The transparent conductive ITO film is made on the front glass. The plural paired display electrodes are formed by a photolithography technology. The metal electrode film of a Cr/Cu/Cr multi-layer is sputtered on these transparent electrodes. The bus electrode is also formed by a photolithography technology. These electrodes are covered with a frit glass layer with a printing technology and then heated at about 600°C to make a transparent dielectric layer. The seal glass layer with a width of about 3 mm is made on the outside of the display area and then pre-heated. An MgO protecting layer is evaporated on the dielectric layer over the display area of inside of the seal layer. The front plate is completed with these processes.

A small hole of a diameter of about 1 mm is made on a corner of the rear plate. The Ag address electrodes are printed and heated. The frit glass is printed on the electrodes in the display area and then

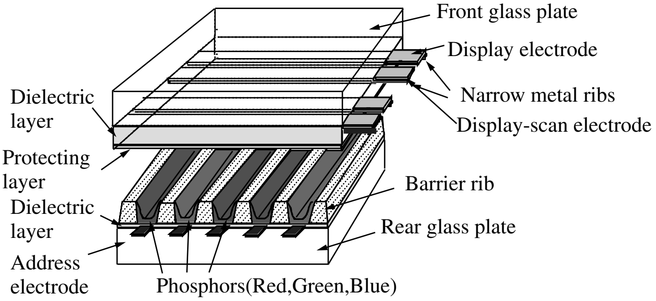


Figure C2.2.24. The practical panel structure of the three-electrode surface discharge PDP with the stripe structure.

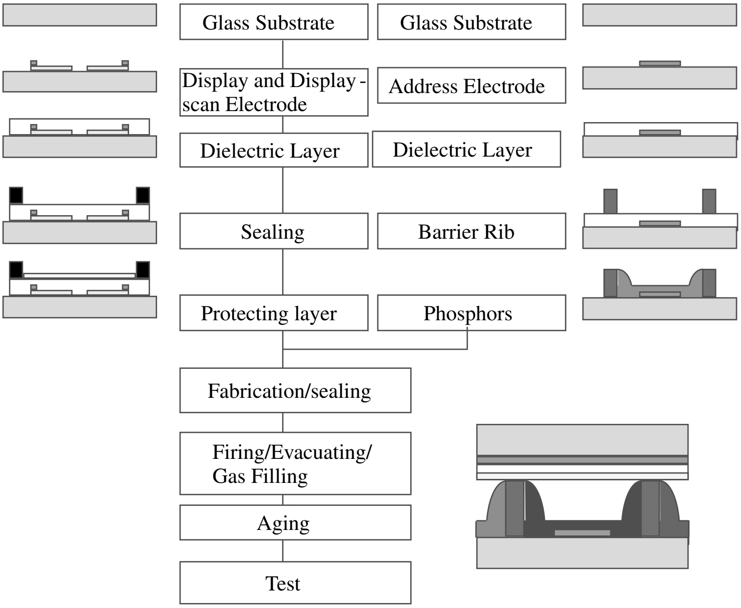


Figure C2.2.25. The essential fabrication process of colour plasma display panel.

heated at about 600°C to make a dielectric layer. The barrier ribs are made by sandblasting the frit glass on both sides of the address electrode. The red, blue and green phosphors are printed inside the channel between the barrier ribs. Each colour phosphor is printed at a same time and then the printing is repeated three times and then dried. The rear plate is completed with these processes.

Both plates are assembled and fixed with clips. The assembled plates are heated to melt the seal layer and the plates are glued resulting in a panel. At the next gas filling process, the panel is connected to an evacuation/gas-filling system through the evacuating tube. After the baking, the discharge gas is introduced. Finally, the PDP is completed after melting off the evacuating tube. The driving pulse is applied to the panel and discharges are ignited in every discharge cell to reduce and make the operating voltage stable.

C2.2.2.5 Improvement of the cell structure

Fujitsu Hitachi Plasma Display (FHP) had reported on the alternate lighting of surfaces (ALIS) technology as shown in figure C2.2.26 [28]. The ALIS method does not have a non-luminous area. Discharge takes place between adjacent display electrodes, instead of scan and sustain electrode pairs one by one. This system, however, does not permit line progressive scanning by upper and lower discharge cells with the use of shared cells. This interlaced scanning can be operated by the drive circuits, which are the same as those in existing PDPs with 480 scanning lines. It does not require special high-speed addressing technology or dual scanning, which requires twice the usual number of ICs. It is possible to double its definition by using the same manufacturing and driving technology as that for the conventional method. In other words, they can apply conventional methods to turn a VGA colour PDP into an SXGA PDP.

The ALIS system itself does not improve luminous efficacy. However, the adoption of the method makes a non-luminous area unnecessary and raises the aperture rate to 65%. As a result, ALIS improved the luminance by 150% of that of the conventional method. By adopting the ALIS system, FHP made a

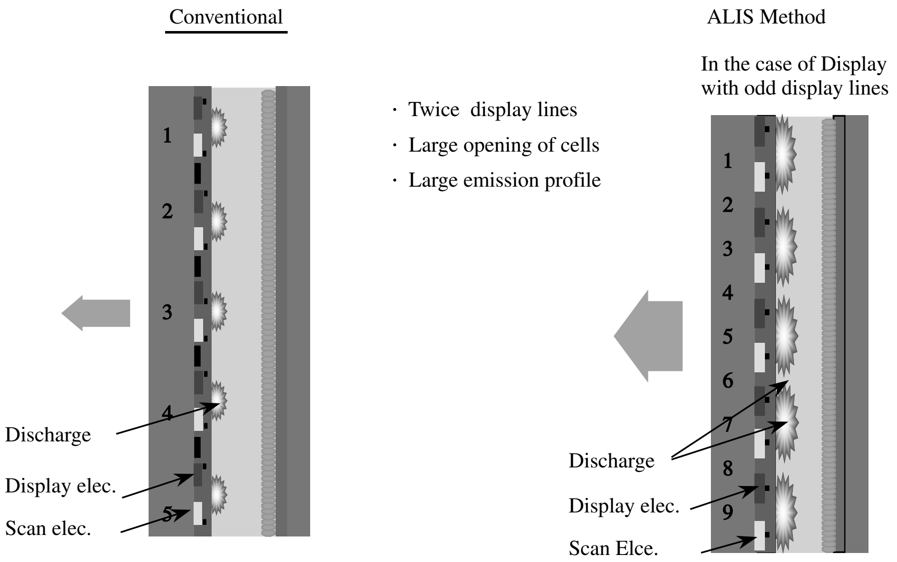

Conventional
1
2
3
4
5
Discharge
Display elec.
Scan elec.
· Twice display lines
· Large opening of cells
· Large emission profile
ALIS Method
In the case of Display
with odd display lines
1
2
3
4
5
6
7
8
9
Discharge
Display elec.
Scan Elce.

Figure C2.2.26. The alternate lighting of surfaces (ALIS) technology developed by FHP.

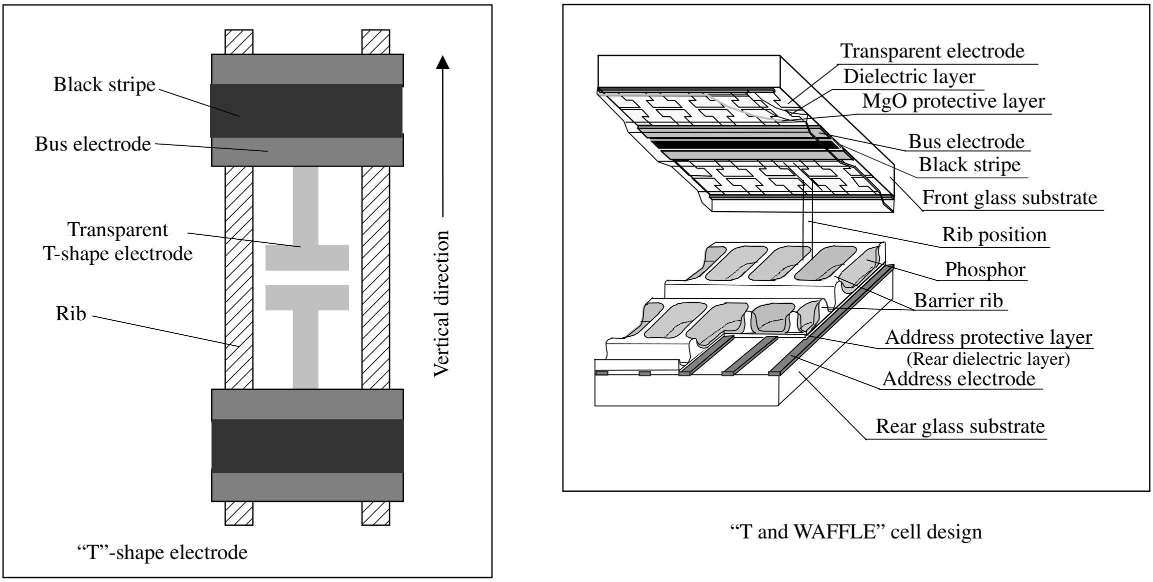
32-in. colour PDP with 852 × 1024 pixels capable of displaying high definition pictures. Figure C2.2.27 shows the 42-in. diagonal HDTV with ALIS technologies.

Pioneer devised PDPs with a high luminous efficacy and a high contrast ratio by adopting a T-shaped electrode structure as shown in figure C2.2.28 [29]. The T-shaped electrode structure produces a favourable effect on luminous efficacy and contrast ratio.

In the conventional system, the ribs on the rear plate side, to which the RGB phosphors are applied, adopted a stripe structure. The waffle structure arranges these ribs in parallel crosses [30]. The waffle rib structure eliminates light leaks vertically, to reproduce sharp image contours. At the same time, the arrangement can widen the per cell area of applied phosphors. The adoption of the T-shaped electrode structure and the waffle rib structure enables a luminance of 560 cd m^{-2}, which is 60% higher than that of the conventional system, raises luminous efficacy by 40% and produces a colour PDP with optimum high resolution.

Method	Alternate lightning surface method
Display size	922 mm (H) × 522 mm (V)
Number of cells	1,024 (H) × 1,024 (V)
Sub-pixel pitch	0.9 mm (H) × 0.51 mm (V)
Colours	16.77 million colours
Luminance	500 cd m^{-2}
Contrast ratio	250 : 1 (dark room)
Power consumption	250 W$_{max.}$

Figure C2.2.27. A 42-in. HDTV with ALIS technologies.


Black stripe
Bus electrode
Transparent T-shape electrode
Rib
Vertical direction
Transparent electrode
Dielectric layer
MgO protective layer
Bus electrode
Black stripe
Front glass substrate
Rib position
Phosphor
Barrier rib
Address protective layer
(Rear dielectric layer)
Address electrode
Rear glass substrate

"T"-shape electrode

"T and WAFFLE" cell design

Figure C2.2.28. Three-electrode surface discharge plasma display with waffle rib and T-shaped electrode structure developed by Pioneer.

C2.2.2.6 Gradation

The display is performed by controlling the two states, ON and OFF, in AC PDP. Changing the number of sustain pulses changes the luminance. The wall charge especially plays an important role to control the ON and OFF states in AC PDP and a sufficiently wide pulse width is needed to get a sufficiently wide operating margin. The ADS method is adopted for operating the AC PDP to meet the requirement and realizes an HDTV image with 256 greyscales. Figure C2.2.29 shows the ADS method [31]. An image is constructed with 60 fields and each field is divided into eight sub-fields. Each of the eight sub-fields has different luminous level and the example of the luminance ratio for the eight sub-fields is 1:2:4:8:16:32:64:128 to realize the 256 greyscales. The luminance level is determined by setting the number of sustain pulses for each sub-field. The 256 greyscales are realized with the combination of the sub-fields in which the ON and OFF states are controlled by depending on the display data.

Each sub-field is divided into two periods, such as address period and display period, as shown in figure C2.2.30. In the address period the discharge is ignited between selected address electrodes and a scan electrode sequentially depending on the display data. And then the wall charges are formed in the selected cells all over the display area and the successive sustain pulses are applied in the display period between all display electrodes. The width of the address pulse is less than 2 μs, and then the addressing speed is sufficiently high to realize the 256 greyscales for HDTV format.

The address period is further divided into the reset step and address step. The reset step is important for operating AC PDP. The wall charges accumulated in the previous display period are eliminated and then the pre-condition is made to ignite the discharge stable in the address step. Although the display period and address step are indispensably important, the reset step is also important for stable addressing, high speed addressing and controlling the contrast.

The ADS method has advantages, such as an easiness in setting the width of sustain pulses, the number of sustain pulses, low power consumption, controllability of the greyscale and stable operation.

The ADS has been improving to realize a high quality display image and stable operation.

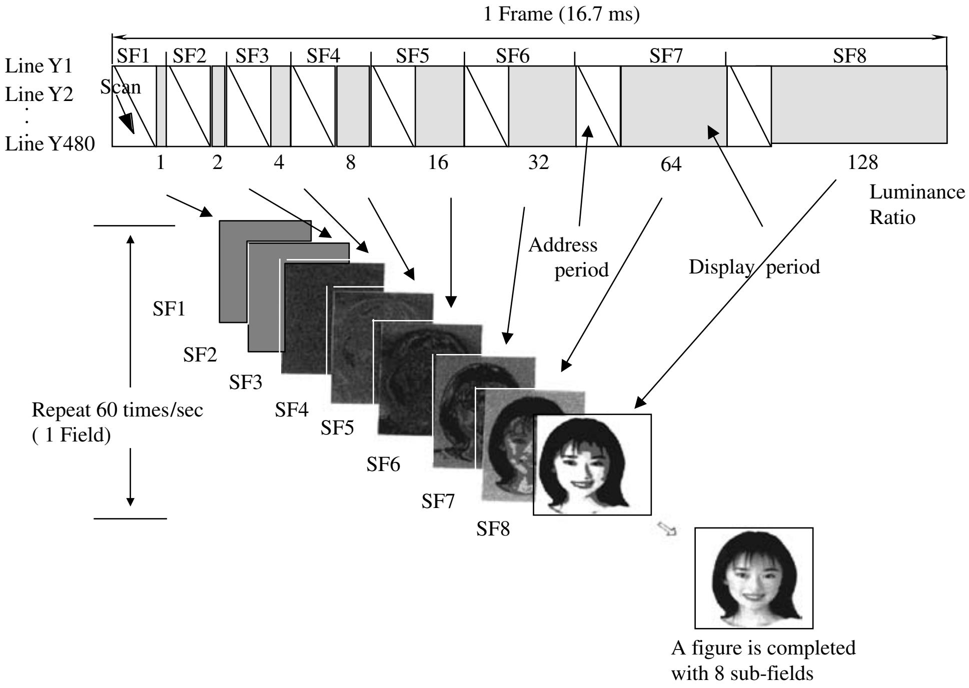

1 Frame (16.7 ms)
SF1
SF2
SF3
SF4
SF5
SF6
SF7
SF8
Line Y1
Line Y2
:
Line Y480
Scan
1
2
4
8
16
32
64
128
Luminance Ratio
Address period
Display period
Repeat 60 times/sec (1 Field)
SF1
SF2
SF3
SF4
SF5
SF6
SF7
SF8
A figure is completed with 8 sub-fields

Figure C2.2.29. The ADS method, address-, display-period separation method, adopted for operating the AC PDP to realize an HDTV image with 256 greyscales.

C2.2.2.7 Improvement of drive systems

Characteristically, many colour PDPs now adopt software-based improvements in addition to the hardware-based improvements such as the improvement of cell structures.

False contour issue was one of the essential issues to degrade the image quality when PDP adopts a driving method for greyscale using the luminous combination of different luminance sub-fields as shown

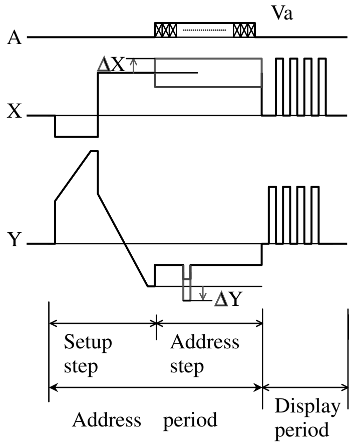

Va
A
ΔX
X
Y
ΔY
Setup step
Address step
Address period
Display period

Figure C2.2.30. Detailed waveform for ADS method. Each sub-field is divided into two periods, such as address period and display period.

in figure C2.2.29. They have always been one of the most serious problems that degraded the display image on PDPs. To diminish the false contours, engineers previously resorted to methods that made false contours as invisible as possible. There are many methods to reduce the false contours, such as error diffusion, dithering, duplicated sub-field method, and so on [32]. With the combination of the methods, the false contour issue is improved as an acceptable level for the television application. Figure C2.2.31 shows the duplicated sub-field method.

Pioneer achieved a high contrast ratio of 560:1 and solved the problem of false contours by adopting a new ADS sub-field method called the Hi-Contrast & Low Energy Address & Reduction of False Contour Sequence (CLEAR) system, to achieve a Hi-Definition progressive display with excellent picture quality [33].

The CLEAR method resolved this problem by preventing, in principle, the generation of false contours by using the luminance accumulation of different luminance sub-fields as shown in figure C2.2.32. Although one TV field is divided into sub-fields, which have different luminance in each the same as the conventional one, the reset period is only once. The discharges are ignited at once in all cells of the panel and form wall charges in the reset step. The wall charges are erased at the address period in the selected sub-fields according to the display data and then eliminate the sustaining discharge in the display period of the sub-field. So the luminance of unselected cells is accumulated from the first sub-field to the sub-field just advanced to the sub-field in which the erase pulses are applied. When the luminance is gradually varied, the light emission pattern of the sub-field does not change largely as in the case of the conventional sub-field method. Then the false contour issue is essentially solved. If the principle described above is applied simply, the grey levels are insufficient to display a beautiful image. To get more grey levels, the number of sustain pulses of the sub-field is changed in each TV field. When a TV field is composed of the m sub-fields and the sustain pulse number is changed between the n TV fields, an $[m \times n + 1]$-step greyscale is obtained. And dither method and error diffusion method are also applied; then, the CLEAR system yields a colour PDP with the same gamma characteristic as that of CRTs. Consequently, a display capability of over 256-step grey levels for each RGB cell can be realized.

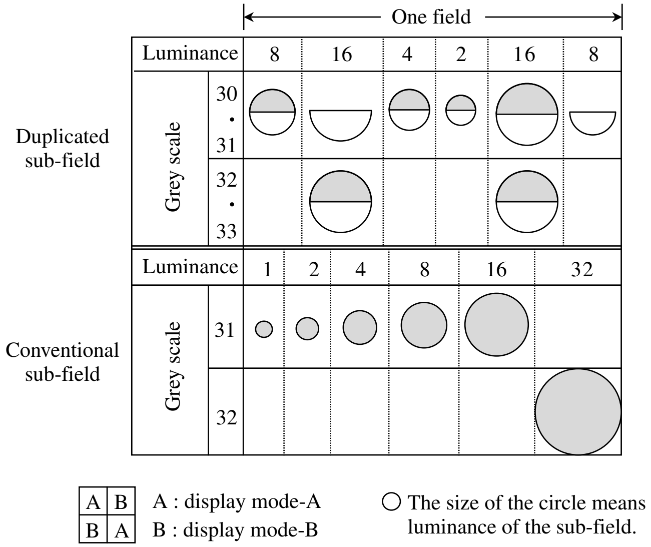


Figure C2.2.31. Suppression of the dynamic false contour by the duplicated sub-field method.

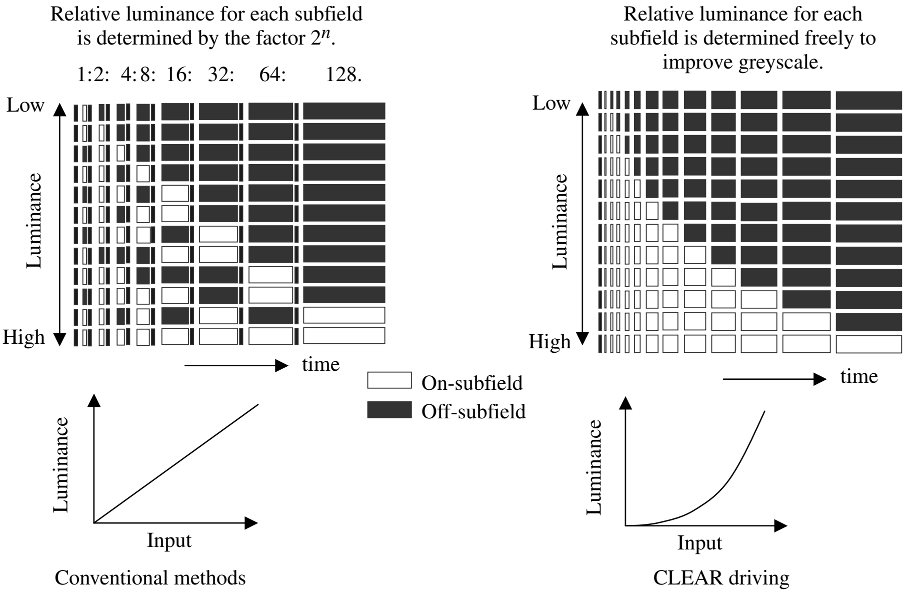

Relative luminance for each subfield is determined by the factor 2^n.
1:2: 4:8: 16: 32: 64: 128.
Low
Luminance
High
time
Luminance
Input
Conventional methods
On-subfield
Off-subfield
Relative luminance for each subfield is determined freely to improve greyscale.
Low
Luminance
High
time
Luminance
Input
CLEAR driving

Figure C2.2.32. The CLEAR method resolved the problem by preventing, in principle, the generation of false contours by using the luminance accumulation of different luminance sub-fields.

C2.2.2.8 Future prospects of colour PDP

One of the latest innovative developments is the delta arrangement PDP with meander barrier ribs reported by Fujitsu Labs [34]. This improves luminance and luminous efficacy by increasing the discharge cell size and also the area of phosphor application as shown in figure C2.2.33. The luminance and luminous efficacy of this structure are about double those of the conventional method.

The next approach to improving performance was the idea of raising the concentration of Xe in the conventional rate of 4 or 5%. It is well known that this method did not attract much attention,

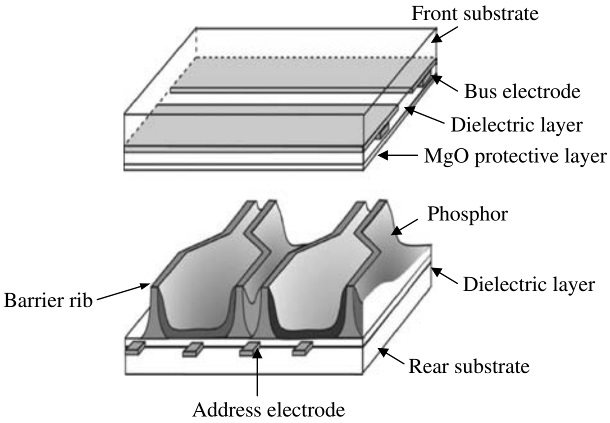

Front substrate
Bus electrode
Dielectric layer
MgO protective layer
Phosphor
Dielectric layer
Barrier rib
Rear substrate
Address electrode

Figure C2.2.33. Delta arrangement PDP with meander barrier ribs. This improves luminance and luminous efficacy by increasing the discharge cell size and also the area of phosphor.

Figure C2.2.34. A 61-in. diagonal PDP (NEC).

because it did not solve the problem of the extra cost for the driver ICs that must handle the additional driving voltage. Nevertheless, it certainly was effective in improving the luminous efficacy and luminance of PDPs.

C2.2.2.9 Conclusion

As we have already explained, plasma television has rapidly penetrated the market since 2001, the first year of the plasma television era.

The production of the large area plasma display was started in 1996 for industrial use and since then a larger market than the expected one has been built. Currently, the big new market of large area TV has been opened for plasma displays and PDP is expected to get a promising position as a key device for the new large area display market from 30 to 80-in. diagonals. Figure C2.2.34 shows the 61-in. diagonal colour PDP.

The next technical issues to make the PDP market larger are reduction of the power consumption, development of highly efficient production processes and circuit technologies for low cost. Especially, increasing the luminous efficacy to reduce the power consumption is the most important issue to develop the ideal TV set with low driving cost. The luminous efficacy is around $1\,\mathrm{lm\,W^{-1}}$ in the current products and the largest limit will be from 2 to $3\,\mathrm{lm\,W^{-1}}$ under the condition on the line of the current technologies. This will be able to make the PDP highly competitive to LCD and CRT. However, to realize the ideal thin and lightweight wall hanging TV having the performance, such as 100 W power consumption, 3 cm thickness and 10 kg weight for 42-in. diagonal PDP, research for the breakthrough in the fields of a discharge system, operating method, panel structure and materials is strongly expected.

References

[1] Piece J T 1954 *Theory and Design of Electron Beams* (Princeton, NJ: van Nostrand)

[2] Boekhorst A and Stolk J 1962 *Television Deflection Systems* (United Kingdom: Cleaver-Hume)

[3] Morrel A M, Law H B, Ramberg E G and Herold E W 1974 *Color Television Picture Tubes* (New York: Academic)

[4] Tannas L E Jr 1985 *Flat-Panel Displays and CRTs* (New York: Van Nostrand Reinhold)

[5] Hawkes P W 2002 *Advances in Imaging and Electron Physics vol 105* (New York: Academic)

[6] Wada Y and Daimon T An electron gun for 76 cm 120-degree 16.9 color TV tubes *SID01 Digest* 1116–1119

[7] Maeda M 2 inch flat CRT *Japan Display83, PD*

[8] Okano N, Maeda M, Saita K and Horiuchi Y Development of ultra-high resolution 17/21″ CRT *SID99 Digest* 254–257

[9] Bergman A H, van den Brink H B, Budzelaar F P M, Engelaar P J, Holtslag A H M, Ijzerman W L, Krijn M P C M, van Lieshout P J G, Notari A and Willemsen O H The fast intelligent tracking (F!T) tube: a CRT without a shadow mask *SID Digest* 1210–1213

[10] Sugawara T and Murakami T Status of glass bulb development for flat and thin CRTs *SID02 Digest* 1218–1221

[11] van der Vaart N C, van Gorkom G G P, Hiddink M G H, Niessen E M J, Rademakerts A J J, Rosink J J W M, Winters R and de Zwart S T A nobel cathode for CRTs based on hopping electron transport *SID02 Digest* 1392–1395

[12] Shinoda T, Wakitani M and Yoshikawa K 1998 High level gray scale for AC plasma display panels using address-display-period-separated sub-field method *Trans. IEICE* 3 **J81-C-2** 349–355 in Japanese

[13] Yoshikawa S, Kanazawa Y, Wakitani M, Shinoda T and Ohtsuka A 1992 Full color AC plasma display with 256 gray scale *Japan Display'92* 605–608

[14] Hirose T, Kariya K, Wakitani M, Ohtsuka A and Shinoda T 1996 Performance features of a 42-in.-diagonal color plasma display *SID 1996 Digest* 279–282

[15] Hoehn H J and Martel R A 1973 Recent developments on three-color plasma panels *IEEE Trans. Electron. Devices* **ED-20** 1078–1081

[16] Kojima T, Toyonaga R, Sakai T, Tajima T, Sega S, Kuriyama T, Koike J and Murakami H 1979 Sixteen-inch gas-discharge display panel with 2-lines-at-a-time driving *Proc. SID* **20** 153–158

[17] Kurita T, Yamamoto T, Takano Y, Ishii K, Koura T, Kokubun H, Majima K, Yamaguchi K and Murakami H 1998 Improvement of picture quality of 40-in.-diagonal HDTV plasma display *Proc. IDW'96* **2** 287–290

[18] Shinoda T and Niinuma A 1984 Logically addressable surface discharge ac plasma display panels with a new write electrode *SID 1984 Digest* 172–175

[19] Bitzer D L and Slottow H G 1966 *AFIPS Conf. Proc.* **29** 541

[20] Takashima K *et al* 1973 Surface discharge type plasma display panel *SID 1973 Digest* 76–77

[21] Dick G W 1974 Single substrate AC plasma display *1974 SID Int. Symp., Dig. Tech. Papers* 124–125

[22] Shinoda T *et al* 1980 Surface discharge color AC-plasma display panels *Late News in Biennial Display Research Conference*

[23] Uchiike H *et al* 1986 Mechanisms of 3-phase driving operation in surface-discharge ac-plasma display panels *1986 Int. Display Res. Conf.* 358–361

[24] Shinoda T, Miyashita Y, Sugimoto Y and Yoshikawa K 1981 Characteristics of surface-discharge color ac-plasma display panels *SID Int. Symp. Digest* 164–165

[25] Dick G W 1985 Three-electrode per pel AC plasma display panel *1985 Int. Display Res. Conf.* 45–50

[26] Dedule M C and Chodil G J 1975 High-efficiency, high-luminance gas-discharge cells for TV display *1975 SID Symp. Digest* 56–57

[27] Shinoda T, Wakitani M, Nanto T, Awaji N and Kanagu S 2000 Development panel structure for a high resolution 21-in.-diagonal color plasma display panel *IEEE Trans. ED* 1 **47** 77–81

[28] Kanazawa Y *et al* 1999 High-resolution interlaced addressing for plasma displays *Proc. SID'99* 154–157

[29] Amemiya K and Nishio T 1997 Improvement of contrast ratio in co-planar structured AC-plasma display panels by confined discharge near the electrode gap *Proc. IDW'97* 523–526

[30] Komaki T, Taniguchi H and Amemiya K 1999 High luminance AC-PDPs with waffle-structured barrier ribs *IDW'99 Digest* 587–590

[31] Shinoda T *United States Patent* 5,541,618

[32] Makino T, Mochizuki A, Tajima M, Ueda T, Ishida K and Kariya K 1995 Improvement of video image quality in plasma display panels by suppressing the unfavourable coloration effect with sufficient grey shades capability *Proc. Asia Display* 381–384

[33] Tokunaga T, Nakamura H, Suzuki M and Saegusa N 1999 Development of new driving method for AC-PDPs *Proc. IDW'99* 787–790

[34] Toyoda O *et al* 1999 A high performance delta arrangement cell PDP with meander barrier ribs *IDW'99 Digest* 599–602

C2.3
Liquid crystal displays

David Coates

C2.3.1 Introduction

Even until the 1960s many eminent scientists regarded liquid crystals as ‘academic curiosities’. This changed in the early 1970s when, from humble beginnings in very simple watch displays, liquid crystal displays advanced into everyday items and became state-of-the-art flat screen colour monitors. However, there are many liquid crystal display types, each with its own set of properties delivering displays optimized for different markets and providing formidable competition to new technologies.

In this chapter the basics of liquid crystal science, how liquid crystal displays are made and a selection of typical display modes will be described.

C2.3.2 Fundamentals of liquid crystal phases

Liquid crystals are a fourth state of matter that exists between a crystal and a liquid. They were first recognized as such by Reinitzer [1] in 1888 and by the early 1920s the liquid crystal phase was widely accepted as a new phase of matter. Display applications had been suggested for liquid crystals since the 1930s, but none were realized because the liquid crystal phase always occurred at an inconveniently high temperature and unless heaters were used in the application the liquid crystal would crystallize. Thus consumer devices based on liquid crystals were impractical. To overcome this problem, research to understand the role of molecular size, shape and polarity on liquid crystal phase formation was carried out with great success [2–4].

All liquid crystal displays are made from liquid crystal phases formed from elongated or rod shaped molecules. While there are other types of liquid crystal, these will not be discussed.

In the solid phase elongated molecules prefer to pack into a crystal lattice similar to the one shown in figure C2.3.1.

When such a crystal is heated, changes in the positional order of the molecules occur. Initially, the attractive forces between the ends of the molecules are overcome by the heat energy (the melting point), a layer-like arrangement of molecules forms and is termed the smectic liquid crystal phase. This phase is very viscous and will not pour easily from a vial. While these molecular layers are more imaginary than real the phase does behave as though it did have layers of molecules that can flow over one another. Currently six smectic variants are recognized; these are differentiated by how the molecules are packed within these layers and whether the molecules are tilted or orthogonal within the layers. At a higher temperature, the attractive forces between the sides of the molecules are partially overcome and the molecules can longitudinally flow past each other, i.e. the smectic layers interdigitate and the nematic liquid crystal phase is formed. This phase is a fluid, usually white (due to light scattering) turbid fluid

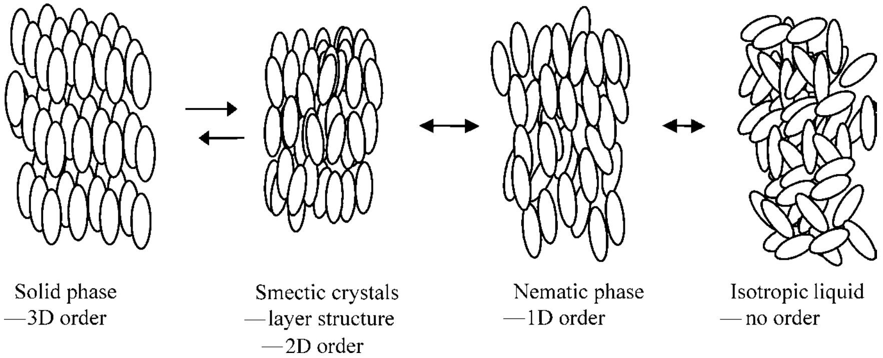


Figure C2.3.1. Schematic drawing of the effect of heat on a solid 3D arrangement of elongated molecules. The solid changes into a layered structure (smectic) then the layers flow into each other forming a nematic liquid crystal phase and finally the isotropic liquid forms. The liquid crystal transitions are reversible on cooling.

that is easily poured from a vial. At a higher temperature (the clearing point) more residual attractive forces are overcome and the molecules become free to move around in a random manner—this is the clear, fluid, isotropic liquid. These changes are thermally reversible except the change to the crystal phase that usually exhibits some supercooling. It is common for a compound composed of elongated molecules to exhibit one or two liquid crystal phases. Some compounds exhibit many more but no compound has been found that exhibits all the liquid crystal phase variants. Not all elongated molecules show a liquid crystal phase as in some compounds the melting point is too high. The order in which these phases occur is very precise; more highly ordered ones occur at a lower temperature than the less ordered ones.

Commercially, the nematic phase (N) is the most important because it gives rise to most liquid crystal displays. Minor display applications have been found for two of the least ordered, most fluid types of smectic phase called the smectic A (S_A) and smectic C (S_C). These have a random distribution of molecules in the 'layers'; in the smectic A phase the molecules are orthogonal to the layers and in the smectic C phase they are tilted within the layers.

C2.3.2.1 Chiral nematic liquid crystals

Another important variant occurs when the molecules contain a chiral centre. In solution such compounds exhibit optical activity and rotate polarized light. When the racemic version of the compound (a mixture of both optical active isomers and thus not optically active) forms a nematic liquid crystal phase, the chiral form exhibits a chiral nematic phase (formerly called a cholesteric phase) whose temperature range is the same as the nematic phase of which it is the direct optically active analogue.

It is imagined that 'sheets' of molecules, arranged as in the nematic phase, are twisted with respect to each other due to the asymmetric force field and steric asymmetry of the chiral molecules. (This is graphically represented in figure C2.3.2.) A line following the direction of the long axes of the molecules within these sheets describes a helix. It exhibits unique optical properties that have been well documented [5]. The sheets of molecules act as a Bragg reflector formed by layers of like refractive index where the helical pitch (P) is related to the distance between layers. In a well-aligned sample with light impinging along the helical axis the familiar Bragg equation (C2.3.1) is generally obeyed:

$$\lambda_0 = \tilde{n} P \cos\theta \tag{C2.3.1}$$

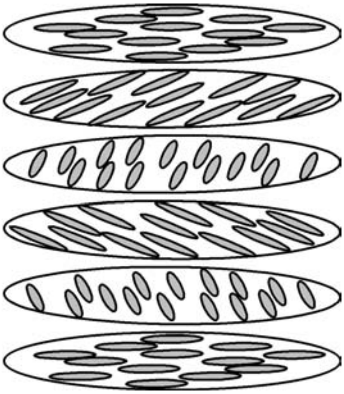


Figure C2.3.2. Schematic drawing of the 'sheets' of molecules arranged in a 'nematic-like' orientation but undergoing a twist from one sheet to the next and forming a helix.

where θ is the angle of view from normal incidence and $\tilde{n}$ is the mean refractive index. Thus as the viewer moves away from the normal, the centre of the reflected wavelength (λ_0) becomes shorter.

Elongated molecules in the liquid crystal phase exhibit two refractive indices (see C2.3.3.1) and thus there is a bandwidth of reflections (equation (C2.3.2)) rather than a single wavelength.

$$\Delta\lambda = \Delta n P. \tag{C2.3.2}$$

The waveband of reflected light is circularly polarized and has the same handedness as the helix; the opposite handedness of circular polarized light is transmitted. Thus, 50% of the light in this waveband is reflected and 50% is transmitted. All other wavelengths are transmitted and rotated.

Any optically active molecule can be added to a nematic liquid crystal and it will transform the nematic phase into a chiral nematic phase. The chiral compound does not have to be liquid crystalline itself, indeed in many applications this is often not the case; the host nematic compound (or mixture) is used to define the general liquid crystal properties.

The origin of the trivial term 'cholesteric', often used for the chiral nematic phase, derives from the types of compound that were first found to exhibit it, i.e. esters of cholesterol; indeed these were the compounds studied by Reinitzer. These two terms are often used interchangeably.

Chiral smectic phases also exist and are derived in the same way as chiral nematic phases. In this case, the tilted smectics (such as smectic C) are the most important as the loss in symmetry created by the chiral centre and the layer arrangement leads to ferro-, ferri- and antiferro-electric properties [6]. Such liquid crystals have been used in display applications. In this case the 'tilt' direction of the molecules from one layer to the next (*interlayer*), which is usually random, now becomes organized and follows a sequential twist tracing a helical path through the layers. Note that the molecules *within* the layers are usually randomly tilted but can be poled with a small electric field to align the tilt direction of the molecules.

The nomenclature to denote a chiral phase is to add an asterisk such as N* (formerly this was Ch) and Sc*.

Chiral compounds are often added in very small amounts to nematic liquid crystal mixtures to provide a slight bias to influence how the molecules move when they relax after being activated by an electric field (see sections C2.3.5.1 and C2.3.9) or retain a specific twist angle in 'Supertwisted nematic displays' (section C2.3.6). Chiral nematic phases themselves, having a short helical pitch, are also used in displays (section C2.3.10).

C2.3.3 Properties of the liquid crystal phase

The properties of liquid crystal displays rely on the physical properties of the elongated molecules from which the liquid crystal phases are derived. Therefore, some knowledge of these properties is helpful to understand how and why liquid crystal displays work.

C2.3.3.1 Refractive index and birefringence (Δn)

The optical properties of liquid crystal displays are largely defined by the birefringence (Δn) of the liquid crystal where ($\Delta n = n_e - n_o$) and n_e is the extraordinary refractive index and n_o is the ordinary refractive index of the nematic phase (figure C2.3.3).

When polarized light is directed at a uniaxial phase (such as an aligned nematic phase in which substantial regions of the sample are arranged with the long axis of the molecules in one direction) the light is split into two rays, the extraordinary and ordinary rays [7]. Due to the asymmetry of the electron cloud on the molecules of the nematic phase each ray travels through the nematic film at a different speed, thus two refractive indices are measured. Changing the magnitude of the electron cloud changes the magnitude of interference that the light experiences and thus the birefringence. Molecules that have a large difference in refractive index between the long and short axes of the molecules have a large birefringence. Some typical liquid crystal compounds and their properties are shown in figure C2.3.4.

The range of Δn is from about 0.06 to 0.27; the ordinary refractive index (n_o) is usually in the range of 1.49–1.52. A few individual compounds have Δn values as high as 0.4 but they tend to be of high molecular weight and this usually confers low solubility in mixtures and thus their practical influence is limited.

C2.3.3.2 Dielectric anisotropy ($\Delta\varepsilon$)

The origins of the anisotropy of the dielectric constants are related to the refractive indices. A molecule can be considered to consist of a series of electric dipole moments composed of permanent and induced dipoles. These dipoles give rise to two dielectric constants along ($\varepsilon_{\text{parallel}}$) and across ($\varepsilon_{\text{perpendicular}}$) the molecular long axis (figure C2.3.3). The difference between these is the dielectric anisotropy ($\Delta\varepsilon = \varepsilon_{\text{parallel}} - \varepsilon_{\text{perpendicular}}$). When an electric field is applied to a thin film of such a material it aligns with the field such that its highest dielectric constant is along the field direction. The magnitude of the difference between the two dielectric constants is related to the voltage required to align the molecule in the field and the relative magnitudes (sign of $\Delta\varepsilon$) define the orientation of the molecule in the electric field.

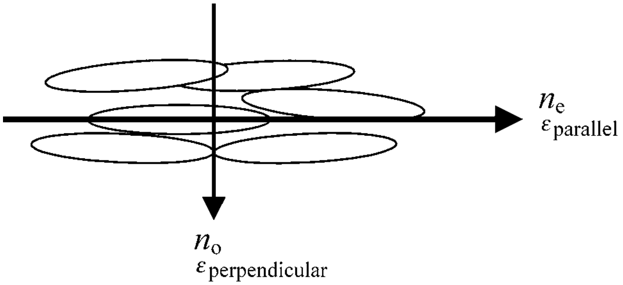

n_e
$\varepsilon_{\text{parallel}}$
n_o
$\varepsilon_{\text{perpendicular}}$

Figure C2.3.3. When arranged in a liquid crystal phase, elongated molecules show two refractive indices and dielectric constants.

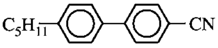
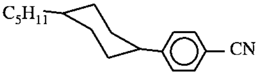
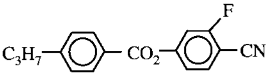
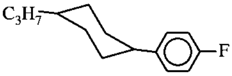
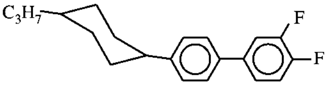
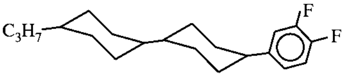
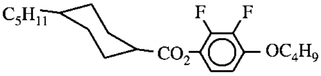
Compound	**Melting point (C-N°C)**	**Clearing point (N-I°C)**	Δn	$\Delta\varepsilon$
	21	35	0.21	+12
	31	55	0.16	+10
	100	46	0.19	+35
	36	–56	0.07	+4
	46	124	0.08	+6
	66	102	0.14	+8
	51	63	0.07	–4.6

Figure C2.3.4. Some typical elongated molecules that exhibit liquid crystal phases and are used in liquid crystal displays.

Both these features can be influenced by suitable chemistry of the molecule. Using highly polar terminal groups such as cyano groups leads to liquid crystal phases that have a high positive dielectric constant (figure C2.3.4) while lower polarity terminal groups (such as fluorine) confer low $\Delta\varepsilon$.

If a polar group is placed across the molecule then it will endow its polarity across the molecule and $\varepsilon_{\text{perpendicular}}$ becomes the larger dielectric constant and the molecule will thus align with its long axis across the electric field. Such materials are said to be of negative $\Delta\varepsilon$ (figure C2.3.4 shows an example). This is fundamentally more difficult to achieve because making the molecule broader forces the molecules to move apart and thus the effect of the attractive forces is reduced, that in turn lowers the liquid crystal phase stability (lowers the transition temperatures). Thus relatively few compounds have a negative $\Delta\varepsilon$, and those that do are not very polar; they are also more viscous than materials of positive $\Delta\varepsilon$. They are however used in liquid crystal display modes on sale today.

In commercial mixtures the $\Delta\varepsilon$ ranges from about -5 to $+35$.

The importance of the $\Delta\varepsilon$ in displays can be seen by its effect on the threshold voltage (V_{th}) of a twisted nematic (TN) display for example as defined by:

$$V_{\text{th}} \propto V_{\text{c}} = \frac{\pi\sqrt{k}}{\sqrt{\Delta\varepsilon\varepsilon_0}} \tag{C2.3.3}$$

where V_{c} is the critical voltage (the voltage at which the first molecules move, figure C2.3.7), k is an elastic constant (typically 10 pN) and ε_0 refers to the dielectric constant of free space.

$\Delta\varepsilon$ is frequency dependent; $\varepsilon_{\text{parallel}}$ in particular becomes significantly smaller as the frequency of the applied field is increased. In liquid crystal phases composed of long molecules, which are inevitably quite viscous, $\varepsilon_{\text{parallel}}$ reduces so much that it becomes smaller than $\varepsilon_{\text{perpendicular}}$ and thus $\Delta\varepsilon$ changes from positive to negative. The point at which this change occurs is the 'cross-over frequency'. Some materials have been optimized such that this occurs at a few kilohertz rather than the usual frequency of tens of kilohertz. This class of materials are the 'two frequency' liquid crystals and can be driven in both directions simply by changing the applied frequency. While this seems very attractive for display use as it would allow fast on and off response times (as the off time can now be driven) it has some major problems: the cross over frequency is very temperature dependent, the materials are inherently viscous and the $\Delta\varepsilon$ values are quite small (typically about $+4$ to $+3$ moving to -1 or -2). They have not been commercialized in displays.

C2.3.3.3 Elastic constants

When an external force distorts a nematic phase, the resistance to this change is described by three curvature elastic constants (k_{11} (splay), k_{22} (twist) and k_{33} (bend)) (figure C2.3.5). These are influenced by the molecular structure. In displays it is often the ratios that are more important rather than the absolute magnitudes; this is especially true for supertwisted nematic displays where the threshold sharpness (steepness of the voltage/transmission curve) is crucially determined by the ratio of k_{33}/k_{11} [8].

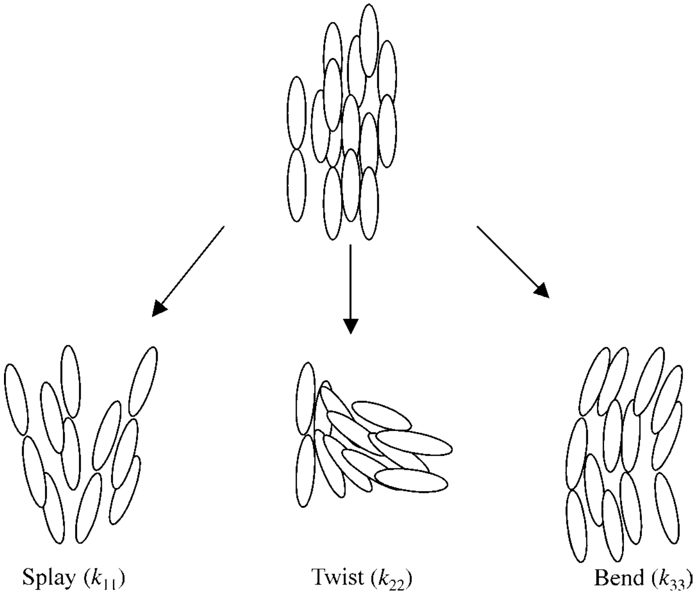


Figure C2.3.5. Elastic constants describe and quantify three deformations of the nematic liquid crystal phase.

C2.3.3.4 Viscosity

The switch-off response speed (or relaxation time τ_d) of liquid crystal displays is dependent on the viscosity (η) of the liquid crystal. Equation (C2.3.4) shows this for a TN display.

$$\tau_d = \frac{\eta}{k\pi^2 d^2} \tag{C2.3.4}$$

where k is an elastic constant and d is the cell gap. The cell gap has a large effect on response times in TN devices (and indeed on most but not all liquid crystal displays). The switch-on time is moderated by the viscosity but is defined mainly by the drive voltage.

There are five viscosity coefficients associated with a nematic phase [9]; often a flow viscosity (η) (measured using a capillary tube) approximates to one of these viscosities that is associated with the elongated molecules flowing past each other with their long axes pointing in one direction. Typical flow viscosities (η) lie between 5 and 100 cSt (mPa s); c.f. water is 1 cSt at 20°C. For predicting or describing the behaviour of supertwisted nematic (STN) displays the flow viscosity is not accurate enough and the more difficult to measure rotational viscosity (γ_1) has to be used. The rotational viscosity is in the range 0.02–0.5 Pa s. The viscosity coefficients have an Arrhenius behaviour similar to isotropic liquids [10].

Compounds with a high $\Delta\varepsilon$ are usually more viscous; thus low voltage operation and fast speed are usually inconsistent. Mixtures having a high nematic to isotropic transition (usually called clearing point) are also more viscous than lower clearing point mixtures. Low birefringence mixtures tend to be less viscous.

C2.3.3.5 Liquid crystal mixtures

No single compound has ever been found that has all the properties to make it useful in a liquid crystal display, thus mixtures are made that contain between 5 and 20 compounds. Some compounds are used to influence the temperature range over which the liquid crystal phase exists and perhaps also eliminate phases that are not required (such as smectic phases). In principle, some components define the optical properties and others the electrical properties and yet others the viscosity. In reality, the compounds perform several of these tasks but can often be regarded as mainly conferring one particular feature.

In general a temperature range of between −20°C and +70 to +80°C is commonly used. For outdoor use, the upper temperature would be increased to about 110°C. For indoor use or in thermostated applications (e.g. wrist watches) the upper range is often lowered to 60°C. It is very difficult to maintain good operation over very wide temperature ranges. As a general rule, the upper operating temperature is about 15°C below the clearing point. Above this point, the optical and electrical properties of the liquid crystal mixture change very rapidly with temperature. The lower operating temperature is usually governed by what response speed is acceptable because it becomes much slower at low temperatures; usually this is in the region of −10 to −20°C. The mixture itself may not crystallize until below −40°C (often quoted as the storage temperature) but it is very unlikely to operate with an acceptable response time at this temperature. The change of response time with temperature, leading to slow response times at low temperatures, is the biggest drawback with liquid crystal displays and opens a niche for other displays that do not suffer from this problem.

C2.3.4 Construction of liquid crystal display cells

Only a brief and simple guide is given here that shows the principles involved and provides some information about the components used. Figure C2.3.6 shows a typical passive drive liquid crystal display cell and its component parts and an active matrix drive colour display.

C2.3.4.1 Substrates

A conventional passive drive monochrome liquid crystal display cell consists of two glass substrates; 0.7 or 1.1 mm thick glass is used although there is a trend to even thinner glass. These are coated with a transparent conducting material; this is usually indium tin oxide (ITO) with a resistance between 10 and $100\,\Omega\,\mathrm{sq}^{-1}$. In the very simplest direct drive cells this would be etched into a seven-segment figure eight, but it could also be a grid of ITO columns on one glass and ITO rows on the other. These rows and columns could be in odd shapes to represent some ‘characters’ in a games display. In an active matrix colour cell ITO is coated over the colour filters on one substrate and deposited on the discrete pixels on the other.

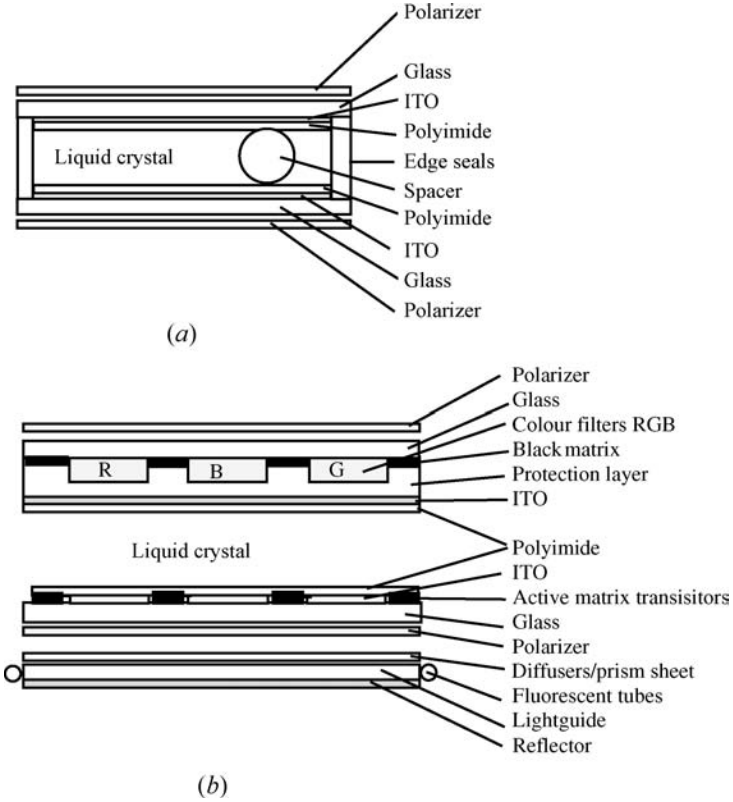

Polarizer
Glass
ITO
Polyimide
Edge seals
Spacer
Polyimide
ITO
Glass
Polarizer
Liquid crystal
(a)
Polarizer
Glass
Colour filters RGB
Black matrix
Protection layer
ITO
R
B
G
Liquid crystal
Polyimide
ITO
Active matrix transisitors
Glass
Polarizer
Diffusers/prism sheet
Fluorescent tubes
Lightguide
Reflector
(b)

Figure C2.3.6. Schematic drawing showing the component parts of a simple monochrome passive drive liquid crystal cell (*a*) and a complex colour active matrix drive liquid crystal cell (*b*).

Very much in fashion at present is the suggestion to use plastic substrates because they are lighter, thinner and flexible. There are several problems preventing this taking place such as the high permeability of plastics to oxygen (typically $>100\,\mathrm{cc\,bar^{-1}\,m^{-2}\,day^{-1}}$) and water (typically $>50\,\mathrm{g\,m^{-2}\,day^{-1}}$). Water and oxygen can have harmful effects on the liquid crystal so the target values for water and oxygen permeability are $10^{-2}\,\mathrm{g\,m^{-2}\,day^{-1}}$ and $10^{-2}\,\mathrm{cc\,bar^{-1}m^{-2}\,day^{-1}}$. To achieve anywhere near these values a barrier layer made from multilayers of evaporated inorganic materials must be used on the plastic. Additionally, the plastic films tend to swell with moisture absorption so a topcoat layer (often a hard polymer film) is also required. After these treatments the plastic is much more expensive than glass! The dimensional stability (expansion coefficient) of plastic films is also high compared to the ITO that would be used as the conducting layer and this leads to cracking of the ITO. If high temperature processes are to be subsequently used then a high T_g for the polymer is required. There are many difficulties to overcome and a market found before plastic displays really become viable.

Amorphous silicon [11] can be deposited onto glass to allow the formation of thin film transistors that allows direct drive to each pixel. These substrates form the basis of active matrix addressed displays (section C2.3.7). Amorphous silicon is formed at low temperatures ($<400°\mathrm{C}$) and is amenable to fabrication in large areas (4–20 inch diagonal) on glass; however, its electron mobility is low and thus the transistors and pixels made from it are large. Such large pixels restrict the size and resolution of the display. Polysilicon has one or two orders of magnitude higher electron mobility and allows very small transistors to be made and thus many small pixels can be used. Unfortunately, it requires high temperatures (1000°C) during its formation and quartz has to be used as the substrate (although local laser annealing methods could allow glass to be used). It is used in small (1.7–4 inch) active matrix addressed high definition displays for viewfinders and also projection displays using transmissive cells. The use of silicon rather than glass as the substrate allows other components to be made on the substrate but because of the high cost it is used in small (<1 inch diagonal) liquid crystal on silicon (LCOS) microdisplays.

C2.3.4.2 Barrier layers

A topcoat or barrier layer is optionally used on top of the ITO layer. This is a thin layer of silicon dioxide whose role is to reduce the possibility of sodium ions from the glass entering the liquid crystal film and of any conducting particle in the display from causing a short across the conductive ITO layers of the cell.

C2.3.4.3 Colour filters

Liquid crystal displays merely act as a shutter to whatever light is shone on them. Usually, white light is used and the familiar black and white displays are formed. To introduce colour, a colour filter is required. In a full colour display, each pixel is divided into three subpixels, red, green and blue. The size of the subpixels is such that when viewed at the intended viewing distance the subpixels cannot be resolved but appear as a uniform colour; this is about $300 \times 100\,\mu\text{m}$ in a monitor display.

Colour filters are made from either dyes or pigments bound in various binders and either printed or coated and then etched into discrete areas. The colours are selected to match the emission lines of the backlight. Because the colours are not monochromatic it is not possible to achieve fully saturated colours (as is also the case with CRTs for example). To display a red colour, two thirds of the light (green and blue) has to be absorbed. Thus, the use of colour filters reduces the light throughput by two thirds.

To prevent light from bypassing the colour filters and reducing the contrast ratio of the display a ‘black matrix’ is used between the pixels (this also reduces the light throughput). To aid alignment of the liquid crystal a smoothing layer is used on top of the colour filters upon which the ITO is coated. ITO is coated on top of the colour filters as they are quite thick insulators (1–3 μm) and would reduce the voltage applied to the liquid crystal too much.

In some reflective liquid crystal displays the relative size of the coloured pixels may differ and an extra non-coloured pixel may be used to obtain extra brightness.

C2.3.4.4 Aligning layer

On top of the barrier layer or topcoat is a thin (30–80 nm) film of polyimide, deposited as a soluble polyamic acid and reacted by heat, *in situ*, to form an insoluble polyimide film. This layer is then rubbed along one direction to form an anisotropic surface upon which the elongated molecules will align to form the required ‘single crystal’ structure. The act of rubbing causes the formation of microgrooves and local heating by friction manipulates the polymer molecules such that any side groups on the polyimide are unidirectionally oriented to give an anisotropic surface. The elongated molecules of the liquid crystal align on the anisotropic surface in the direction of rubbing that creates an accurate and directional pretilt (figure C2.3.7).

There are many polyimides whose main differences are usually the pretilt angle they confer on the liquid crystal. Typical pretilt angles are 1–9°. Some polyimides give 90° tilt angles. Note that the pretilt angle decreases with increasing temperature.

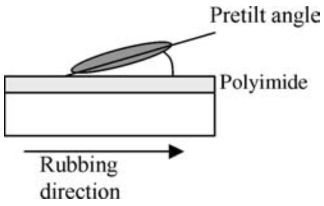


Figure C2.3.7. When a polymer surface is rubbed it aligns the molecules of the nematic liquid crystal phase in one direction with a precise pretilt angle.

More recently some polyimides have been irradiated with polarized short wavelength UV light that anisotropically destroys some polyimide chains such that what is left aligns the liquid crystal. An alternative is to use a polymer layer that is also irradiated with UV but during this process reactive bonds are further polymerized such that an isotropic polymer surface is formed [12].

Early TN displays used obliquely evaporated SiO_x to give low pretilt angles. However, this process was not amenable to mass production. This is now revisited but using atomic-beam deposited inorganic materials [13].

C2.3.4.5 Seal ring

A seal ring is formed on one glass by either screen-printing or syringe dispenser deposition of a UV curable or UV plus heat curable glue. It must provide a strong bond yet resist the outside environment to which it is exposed and not contaminate the liquid crystal. Very few glues can do this.

C2.3.4.6 Cell gap spacing

Spacers made from either glass fibre or plastic spheres are deposited onto the other substrate such that there are about 50–200 per mm^2. In displays that use a silicon substrate they can be formed as pillars on the silicon and are thus integral with the substrate. Spacers define the cell gap; usually this is in the region of 4–6 μm with a tolerance of about $\pm 0.1\ \mu m$.

C.2.3.4.7 Cell assembly

The top glass is placed on top of the bottom glass and pressed down then the glue is cured. The seal ring has one opening through which the air can be evacuated and liquid crystal passed in under the influence of the vacuum in the cell. The filling hole is then sealed with a UV curable glue.

C2.3.4.8 Polarizers

In most applications, film polarizers are glued to the outside of the glass plates. These typically transmit and linearly polarize 40–44% of unpolarized light or about 90% of linearly polarized light with high polarizing efficiency. The active light-absorbing component is usually iodine anisotropically absorbed into stretched polyvinyl alcohol; for high durability (outdoor use) polarizers a dichroic dye is used. A range of antireflection, antiglare, anti-smudge and hardcoat treatments is often put onto the outside of the polarizer film. For polarizers used in reflective displays an aluminium or aluminium/silver reflector is added.

C2.3.4.9 Optical films

In displays that require polarized light the light passing through the liquid crystal at wide angles of incidence often becomes elliptically polarized by the liquid crystal. This elliptically polarized light gives rise to a range of unwanted artefacts (colour changes and contrast reduction and inversion) when it meets the second linear polarizer. Optical films made from stretched polyvinyl alcohol or polycarbonate sheets placed between the cell and the polarizers are used to convert this elliptically polarized light back to linearly polarized light. In this way, these 'compensation' films create a wider viewing angle.

C2.3.4.10 Lighting

Liquid crystal displays do not emit light, they modulate it, and thus light must be obtained from an external source. The commonest light source for PC monitors, etc is a backlight (figure C2.3.6). A very thin (2–8 mm diameter) three phosphor fluorescent tube on one or two sides of the display produces the light. These tubes are coupled to a plastic light guide that has a reflective coating on one side and a prism sheet and diffuser next to the display on the other side. This spreads the light uniformly over the display area.

For some uses, as in automobiles, the tubes are placed beneath the display and may be bent into U shapes.

As most liquid crystal displays absorb most of the light the backlight has to be very bright. For portable computers with a colour display a display brightness of 70 cd m^{-2} is required, and as only 3–6% of the light is transmitted the backlight must emit up to 2000 cd m^{-2}. To compete with a CRT with a screen brightness of 400 cd m^{-2} the backlight must emit 2700 cd m^{-2} for black and white and 13 000 cd m^{-2} for colour displays [14].

Simple passive drive displays often use a small sidelight and the rear polarizer of the display has an aluminium reflector coated onto it. In some cases the reflector may not cover all the area such that a backlight can also be used; these are transflective displays in which the display can be used in either reflected or transmitted light.

C2.3.4.11 Brightness enhancement films

In displays using polarizers most light is lost. This is a major energy drain for a portable computer battery for example. Reflective polarizers have been developed to help convert non-polarized light to polarized light yet without the usual absorption losses. There are two types, a multi-layer laminated film [15, 16] and a polymerized chiral nematic based reflective polarizer [17, 18]. Both sit between the backlight and the first polarizer. Both transmit linear polarized light and reflect back into the backlight any light not transmitted so it can be recycled and converted again into linear polarized light that will be allowed to pass through the film. Perfect reflection by these complex structures is very difficult to achieve and this leads to some colouration and reduction in efficiency at wide viewing angles. In this way up to about 80% of the light is converted to polarized light (rather than less than 50%).

Another film that directs light from wide angles to the centre is also available (BEF film from the 3M Corporation); this 'prism film' is composed of many clear triangular lenses and can increase the brightness but as it simply directs light away from outside these regions become darker.

C2.3.4.12 Electrical connection

Driving of the cell is by connecting the ITO pads to the outside drivers with either anisotropically conducting rubber strips that are pressed between a many pin connector and the corresponding pattern of ITO electrodes or a flexible polymer connector that is glued to the ITO using conducting glue. In more complex displays, drive chips are bonded to the perimeter of the display (chip on glass technology) directly on top of the ITO tracks and the chips are then connected via fewer flexible connectors to the main drive boards.

C2.3.5 The development of liquid crystal displays

The first compound (4-methoxy benzylidene 4′-buylaniline—MBBA) to show a nematic phase at room temperature was made in 1969 by Kelker *et al* [19] and while it represented a milestone, it was unstable to moisture. This compound and some of its homologues were used in some early 'dynamic scattering displays' [20]. While the dynamic scattering display had poor viewing characteristics, a short life and drew a lot of electrical current it was used in early watches and simple calculators. These first displays illustrated what liquid crystals could do; consumers liked the light weight and low power consumption compared to the LED displays of that time.

Dynamic scattering displays were replaced by the TN display [21] that utilized light polarizing films. While such displays were not perfect, their viewing angle, response times, low power consumption, high visual contrast, versatility for use in many applications and low cost have proven a difficult combination to beat. TN displays have dominated the flat panel displays market ever since. This was helped by the discovery by Gray *et al* [22] of the family of 4-alkyl-4′-cyanobiphenyls (figure C2.3.4—first compound in the table). These compounds exhibited a nematic liquid crystal phase at room temperature, they had the correct dielectric properties for use in TN displays and were also stable to light and heat. This was regarded as the start of the liquid crystal display industry.

Hulme *et al* [23] developed these components into eutectic mixtures and later incorporated 4-alkoxy-4′-cyanobiphenyls, 4-alkyl-4″-cyano-*p*-terphenyls and a variety of ester based liquid crystals [24] to widen the nematic temperature range and improve the optical performance. Mixing together such materials depressed the melting point while averaging the nematic to isotropic temperatures and thus widening the temperature range over which the nematic phase was exhibited. These simple mixtures allowed useful display properties with wide nematic phase temperature ranges.

Other analogous compounds such as trans-4-alkyl-4′-cyanophenyl cyclohexanes [25] and 4-alkyl-4′-cyano- 2-phenylpyrimidines [26] were then made and used in competing commercial nematic liquid crystal mixtures.

Thus, during this short period in the early 1970s a robust new device mode was found and liquid crystal materials that would allow its use at convenient temperatures together with polarizer films to allow the display to be visible came together.

C2.3.5.1 The twisted nematic device

As the name suggests, TN displays [21] contain a nematic liquid crystal whose molecules are 'twisted' between the substrates (figure C2.3.8).

A 90° twist is induced in the nematic phase by the two glass substrates, each carrying a layer of rubbed polyimide whose rubbing directions are twisted at 90° relative to each other. The nematic phase behaves in an elastic manner and twists to accommodate the fixing of its boundary molecules onto a surface. Additionally, a small amount of chiral dopant is added to the liquid crystal to convert it to a very long helical pitch chiral nematic phase and aid twisting in the desired direction. It also aids the unidirectional reformation of the twist after switching.

Polarized light is provided by the first polarizer, this light is then passed through and twisted by the liquid crystal until it emerges through the second polarizer set at 90° to the first polarizer—the cell appears bright. When an electric field is applied to the cell, the liquid crystal is oriented such that the light is no longer twisted and becomes absorbed by the second polarizer and the cell appears black.

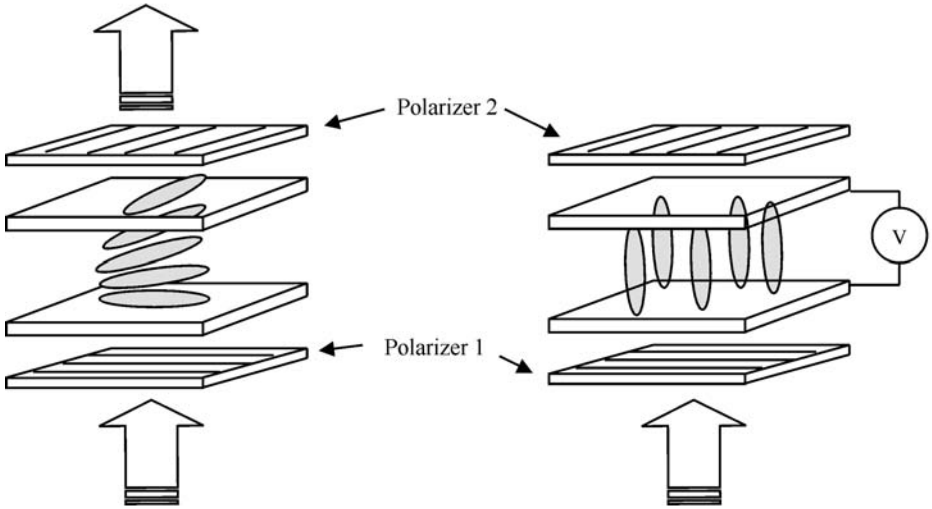


Figure C2.3.8. Operation of a transmissive TN mode device. The off state allows light to pass and is bright, while the on state is dark.

Switching off the field allows the liquid crystal to relax back to its original twisted structure and the display becomes bright again. This relaxation takes between 10 and 50 ms.

The position of the polarizers relative to each other gives rise to two general possibilities assuming that the first polarizer has its transmission axis parallel to the alignment rubbing direction. If the polarizers are crossed (as in figure C2.3.8) then the light that is twisted by the liquid crystal passes through the exit polarizer and the device appears bright when no field is applied; this is the *normally white mode*. The polarizers can also be set with their polarization axis perpendicular to the rubbing direction; this gives better viewing angles for the display. However, if the second or exit polarizer (sometimes called the analyser) is parallel to the first one then the twisted light is absorbed; this is the *normally black mode.*

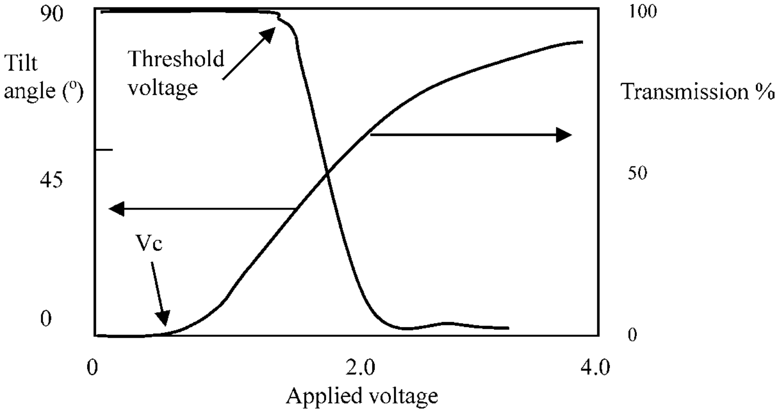


Figure C2.3.9. An applied field causes the molecules in the middle of a TN cell to tilt (mid-layer tilt) at the critical voltage V_C; this eventually causes a change in transmission of polarized light that is seen at the threshold voltage.

This twisting or rotation of the polarization axis of polarized light described above can occur only if the condition in equation (C2.3.5) is fulfilled (Maugiun limit).

$$\Delta nd \gg \lambda \text{ (Mauguin limit)} \tag{C2.3.5}$$

where d is the cell gap, Δn is the birefringence of the liquid crystal and λ is the wavelength of incident light.

To achieve this condition, large cell gaps would be required that would make response times very long (equation (C2.3.4)).

However, Gooch and Tarry [27] found that the transmitted light for a 90° TN cell with parallel polarizers is given by:

$$T(u) = \frac{1}{2}\frac{\sin^2(\pi/2\sqrt{1+u^2})}{1+u^2} \tag{C2.3.6}$$

when $T(u)$ is plotted against u, a series of transmission minima are found that occur when u is $\sqrt{3}$, $\sqrt{15}$, $\sqrt{35}$; these are known as the first, second and third minimum values. Under these conditions the polarized light is rotated by $\pi/2$ without significant deviation. $u = 2d\Delta n/\lambda$ for a 90° TN device and $T(u)$ is the transmitted light intensity.

For the usual normally white mode device the intensity maxima occur for $u = \sqrt{4m^2 - 1}$ (where m is an integer); these correspond to the minima for the normally black mode. These conditions are much more amenable to display applications and reduce to values of $d\Delta n = 0.48$, 1.09 and 1.68 (for $\lambda = 550\,\text{nm}$).

The accuracy of hitting these values becomes less critical as the series is increased. The 'second minimum' devices ($\Delta nd = 1.09$) suit the originally used high birefringence alkyl cyanobiphenyl mixtures and they can have larger cell gap and cell gap tolerance. Unfortunately, due to the larger cell gap second minimum devices have slower response times and a narrower viewing angle than first minimum devices. The first minimum device ($\Delta nd = 0.48$) requires a smaller, more difficult to make and tighter tolerance cell gap, but gives better viewing angles [28]; it is widely used in applications from petrol pump displays to computer monitors etc.

Within these two devices most TN displays are normally white mode because it offers an achromatic and better black state giving higher contrast ratios, wider viewing angle and less colour change with angle. The normally black mode is often used in automotive displays.

C2.3.5.2 Passive matrix addressing of twisted nematic displays

A constraint in addressing schemes is that at all times a balanced dc drive scheme must be used to avoid degradation of the liquid crystal caused by electrolytic effects. Additionally, the liquid crystal is too slow to react to the waveform of the electric field thus it responds to the overall energy which is proportional to the square of the rms voltage averaged over time.

Simple alpha-numeric displays (usually with 3½ digits having 23 segments plus a decimal point) can have a discrete 'wire' or ITO track to each segment; this is known as direct drive. However, as the number of segments increases it soon becomes impractical to make a discrete contact to each segment because there is not enough room around the glass to make contact to so many ITO tracks. A better way of addressing the more complex displays is required.

A solution to the problem of addressing more segments is to use lines (or rows, m) of ITO on one glass plate and at 90° to them on the other glass plate more lines (columns, n) of ITO. At the intersection of a row and column an electrical field can be applied via the row and column electrodes. Each cross-over point is a picture element often called a 'pixel'. This technique allows many more pixels to be addressed using fewer electrical connections, i.e. $m+n$ connections to address $m \times n$ pixels.

Consider the voltage at each pixel. The liquid crystal at a pixel in a normally white TN display will appear ‘white’ (off) when a voltage below V_{off} is applied and it will appear black (on) if the addressing voltage is above the V_{on} voltage (figure C2.3.9). The voltage supplied (V_{p}) to any intersection or pixel will be the sum of the voltages supplied to a pixel defined by the row (V_{r}) and column (V_{c}) voltages. The concept used is to apply a voltage below V_{off} to the columns (so that on its own it has no effect) and then scan sequentially another voltage called the data or video or signal voltage across the rows such when $V_{\mathrm{c}} + V_{\mathrm{r}} > V_{\mathrm{on}}$ the pixel is driven on, but if less than V_{off} the pixel remains off. Alt and Pleshko [29] described how such a scheme could be used. The critical parameter is the difference between V_{off} and V_{on} for the liquid crystal device as this defined the number of rows (N) that could be addressed with optimal contrast. This is defined as:

$$N = \frac{(V_{\mathrm{on}}/V_{\mathrm{off}})^2 + 1}{(V_{\mathrm{on}}/V_{\mathrm{off}})^2 - 1} = N_{\mathrm{max}}. \tag{C2.3.7}$$

Thus to allow many rows of pixels to be addressed a liquid crystal with a steep threshold voltage/transmission is required. The original liquid crystal mixtures for direct drive TN displays did not do this very well.

Figure C2.3.9 shows the relationship between applied voltage and transmittance of the TN cell plus a curve that shows how the tilt angle of the molecules in the middle of the cell changes with voltage. It is clear that the molecules of the nematic phase move at a lower voltage (critical voltage) than what is observed as a threshold voltage for the optical effect that has a much steeper curve. This indicates that the twisted structure can be distorted quite a lot before there is a major effect on the transmission but then the transmission change is rapid. To improve the number of lines that can be addressed this change in transmission versus voltage curve had to be improved.

It was found that by encouraging the formation of a low temperature smectic phase the elastic constants in the nematic phase were altered (due to pre-transitional effects). This gave a steeper threshold curve. Thus, a concept for making ‘multiplex’ addressed nematic mixtures was developed [30]. Ongoing improvements to give lower viscosity (faster switch-on and switch-off times) and lower voltage operation (longer battery life) have made these devices very popular for use in simple games displays, personal digital assistants (PDAs) and calculators.

The trade-off in these displays is between the number of rows that can be multiplex driven and the viewing angle and contrast. The optical performance is not as good as in direct drive TN displays.

C2.3.5.3 Other twisted nematic devices

Many variants of the basic TN device have appeared; two of these are noteworthy.

Hybrid twisted nematic

In recent years, a new breed of TN displays has arisen in which the twist angle is a little more than 90°. This change is based on STN technology (section C2.3.6) in which 180° or more twist is used to improve the threshold steepness and thus give a more multiplexable display. It is claimed that this mode gives a slight but worthwhile improvement in the contrast and performance of TN displays.

Mixed mode twisted nematic

For reflective TN displays (used in many games displays), it has been found that while the polarizers can be retained crossed at 90° the axis of the first polarizer need not be set parallel to the nematic director (along the rubbing direction) [31, 32]. Thus the polarized light from the first polarizer propagates by two modes through the TN cell with the result that in reflection a better black state, independent of wavelength, is produced.

C2.3.6 Supertwisted nematic displays

Multiplex addressed TN displays allow up to about 32 rows of information to be driven; beyond that the optical properties begin to suffer seriously. Liquid crystals for use in multiplex addressed displays with even steeper switching characteristics could not be found, thus a new pathway to more complex displays was sought.

Independently it was shown by Scheffer *et al* [33] and Waters *et al* [34] that by using a larger twist angle in a TN device the threshold voltage curve could be made very sharp indeed. These devices were referred to as STN displays. Various twist angles between 180 and 270° were used; for most uses the 240° twist type has the best compromise of threshold steepness and optical contrast. With ideal liquid crystal materials and with the appropriate cell parameters as many as 240 lines of information can be addressed (often referred to as the duty ratio, in this case 1/240). The display is optimized to act as an electrically switchable light retarder; the on and off states are thus coloured.

The liquid crystal 'twist' is achieved by the addition of chiral dopants and orientation of the rubbing direction on the substrates. New chiral dopants that had high HTP values and very little change in pitch length with temperature had to be made [35].

The steeper transmission/voltage curve can be rationalized with reference to what happens to the molecules in the middle of the cell and to figure C2.3.9. In a stack of molecules aligned in one direction the molecules in the middle of the stack are constrained by their neighbours. If the structure is twisted then the influence of the neighbouring molecules to maintain these positions is weaker. Thus under an applied field the middle molecules will be able to respond more easily to an electric field and the mid layer tilt angle will change more readily to changes in voltage, i.e. it will give a steeper voltage/transmission curve. This is shown in figure C2.3.10 for different twist angles. Above 270° of twist the structure is unstable. The ratio of cell gap (d) and liquid crystal pitch (p) are important and by changes to the surface pre-tilt angle, helical twist and elastic and dielectric constants the window over which stable STN

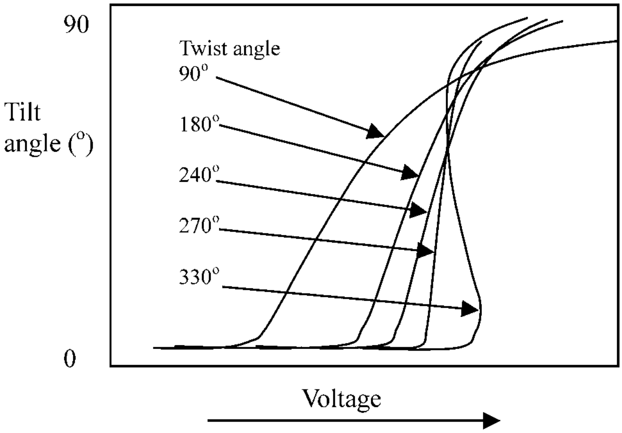


Figure C2.3.10. The voltage/transmission curve and mid-plane tilt angle for STN displays with various angles of twist shows that the steepness of the curve increases as the twist angle becomes larger. After 270° bistability and other unwanted artefacts occur.

operation can be maintained is influenced. This is often referred to as the d/p window. As all these properties vary with temperature the device has a limited temperature range; normally this is designed to be 0–40°C. Because of the extreme sensitivity of the stability of the STN display to cell gap, polished glass and very good cell gap control are required.

To achieve the best contrast ratio the polarizer orientation has to be optimized but (due to the birefringence effects of the liquid crystal) this gives a variety of colour contrasts rather than black and white. For example, with a 270° twist a yellow background with dark blue 'on' state (yellow mode) or a dark purple blue background with white 'on' state (blue mode) are found. Consumers did not like these colours, but it was soon found [36] that a second cell joined to the output of the first cell but with opposite helical twist compensated the light path and changed the light back from elliptical to linear polarization. This gave black and white double STN (DSTN) displays with better contrast and viewing angle. While an advance, this concept was very expensive and added too much weight. However, a compromise solution [37] was found by using retardation films (stretched anisotropic polymers) located between the polarizers and the liquid crystal cell instead of a second cell (FSTN). This works well over a limited temperature range as the temperature dependence of the optical properties of the liquid crystal are different to those of the plastic film.

By substituting an *n*-alkenyl terminal chain for the usual *n*-alkyl chain in typical liquid crystal molecules (figure C2.3.4) the threshold steepness can also be increased due to a higher elastic constant ratio [38]. New 'active' addressing schemes have been tried to improve the speed and contrast ratios [39].

STN displays have modest viewing angles and long response times compared to TN displays; to improve the response times low viscosity liquid crystals are sought and remain an active research topic allied to the search for compounds with large elastic constant ratios [8] and lower voltage operation.

The first STN devices appeared in 1985; they were developed into displays suitable for lap top computer screens by 1987. While used in displays for copying machines and instruments, the major use for STN is now in mobile phones (over 300 million were made in 2002) and PDAs. The main advantage of the STN display is that it allows reasonably complex displays at a competitive (inexpensive) price. It struggles to give adequate performance for portable computers because there are too many rows of pixels to address and this deteriorates the optical contrast and viewing angle too much; it is also too slow.

C2.3.6.1 Colour STN

As information services provide more data the requirement for colour increases. These must be low power for mobile use thus reflective and passive drive displays such as STN become good candidates. A version of the STN display that relies on birefringence colours (STN-ECB effect: [40]) offers better brightness than those with colour filters but is limited to about four colours. Conventional STN displays with two polarizers suffer from low brightness and parallax. Thus, the usual route chosen is a one polarizer STN display with colour filters [41]. These displays require two retardation films and the angles of these relative to each other and the liquid crystal and polarizers have to be optimized. A full colour display with 640 × 480 pixels with a duty ratio of 1/240 gave 15% reflectance and a contrast of 14:1. The response time of 250 ms is much slower than a direct drive TN device.

C2.3.7 Active matrix addressed liquid crystal displays

The search for a method to make complex displays that would rival CRTs was also being followed using a totally different pathway that concentrated on being able to drive directly the TN display and retain its good optical performance.

It had been recognized since the 1970s that a good way to overcome the scanning difficulties of TN displays was to provide a non-linear element at the pixel thus producing a steep threshold characteristic

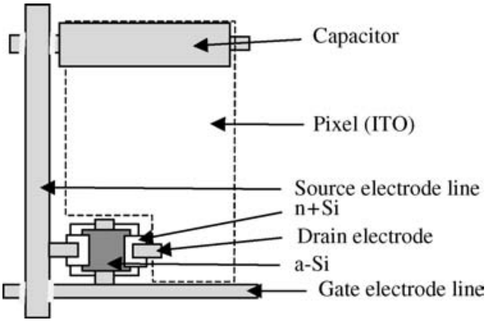


Figure C2.3.11. Structure of a typical transistor that is located within the corner of each pixel.

for each pixel. This concept of 'active matrix addressing' was shown to work [42] using CdSe as the transistor to address small complex black and white displays.

The term active matrix refers to an addressing technique for TN devices (AMTN). TFT/TN is a specific term that refers to the use of transistors for driving a TN display but both tend to be used interchangeably. There are various methods to achieve active matrix addressing [43]. Because of its ease of deposition and overall properties amorphous silicon [11] is usually used. Polysilicon is reserved for small high resolution displays. These transistors act as a valve allowing charge to flow to a drain or pixel when a signal is applied to the gate electrode of the transistor (figure C2.3.11). Metal–insulator–metal devices [44] have been used and give optical performance almost as good as amorphous silicon devices but the fabrication cost is much less. As the number of pixels increases there is an increased chance of some not working, thus redundancy is often built in by having two transistors per pixel.

The liquid crystal at each pixel acts as a capacitor being driven by the transistor. Compared to a multiplex TN device a shallow or less steep threshold/transmission curve for the switch ON is required so that intermediate states (grey levels) can be easily realized by changing the voltage on the pixel. After a field has been applied to the liquid crystal, the liquid crystal has to retain that charge and thus stay in the defined state until the next signal is given to it in the next frame a few tens of milliseconds later. Liquid crystals of the early 1980s did not do this as they contained cyano groups that complexed with ions that lowered the resistivity to typically $10^{10}\,\Omega\,\mathrm{cm}$; this allowed the charge to dissipate and led to poor contrast. This problem was solved using fluorine derivatives (such as F, CF_3, CHF_2 and OCF_3) as the polar group does not complex with ions and they are very stable. Unfortunately, fluorine compounds have poor liquid crystal forming properties so the core of the molecule also had to be improved. The molecules had to allow high resistivity ($10^{13}\,\Omega\,\mathrm{cm}$), low birefringence ($<0.08$ for use in the first minimum mode) and be very fluid ($\eta < 20\,\mathrm{cSt}$) to allow fast response speeds. These requirements were largely fulfilled in the 1990s [45]; some typical fluorinated compounds used in AMTN are shown in figure C2.3.4.

The optics of these displays is essentially that of a standard direct drive TN display, they use first minimum TN technology; only the addressing technique is different. Active matrix displays invariably have an RGB (red, green, blue) colour filter (as in figure C2.3.6) and many pixels (usually XGA resolution 1024 × 768 pixels) equally divided between red, green and blue pixels. Each colour can have up to 16 grey levels (giving 4096 colours). The liquid crystal is very fluid and allows response times suitable for video applications (< 30 ms). Wide viewing angle films are obligatory in the larger displays. XGA 14 inch screens suitable for PC monitors are now widely available. The first AMTN televisions (about 15 inch diagonal) appeared in 2001 and now 20 inch displays are available. Demonstration displays with WXGA (1280 × 768 pixels) in 40 inch diagonal panels are shown in exhibitions.

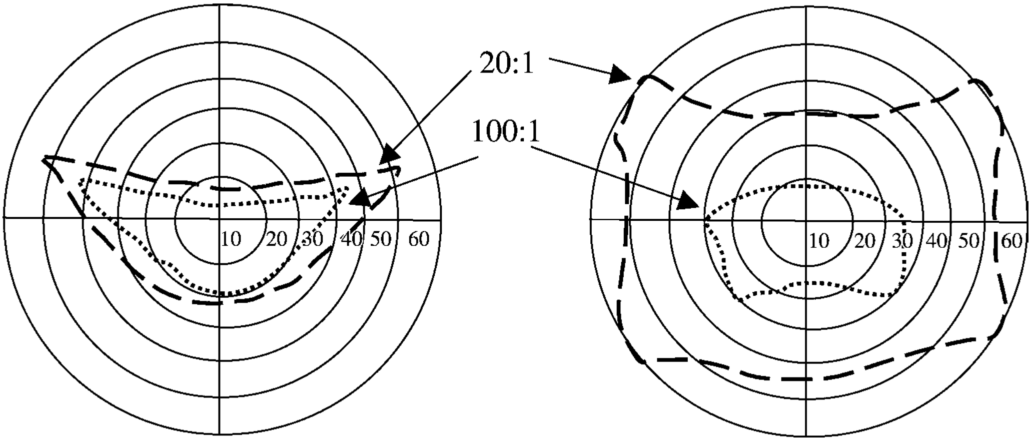


Figure C2.3.12. Iso-contrast plots (100:1 and 20:1) of a TN display and the wider iso-contrast plots of a film compensated TN display.

C2.3.7.1 Wide viewing angle twisted nematic displays

As display screen sizes become larger it is more important to have a good viewing angle to prevent poor images being seen especially in the corners of the display. The contrast ratio of a TN device is high (>200:1) at normal incidence but this falls away to 10:1 and is far from symmetrical (figure C2.3.12). At wider angles colour change and contrast inversion (black becomes white) also occur. While some display modes that compete with TN are better than this, the TN device can be improved. Two techniques are typical.

Optical compensation films

Very good optical compensation films are available that can correct the optical distortion that occurs to linearly polarized light when it passes through a TN structure at wide angles. One such film is made from a discotic liquid crystal polymer [46]. This type of liquid crystal is optically negative and thus opposite to the optically positive nematic phase. It can thus, in principle and practice, be used to compensate the imperfections arising from light passing through a film of nematic material at wide angles. An isocontrast plot of a typical TN device is shown in figure C2.3.12 together with the same cell using a discotic film compensation foil; there is a much wider viewing angle for the 20:1 contrast ratio.

Other films using reactive liquid crystals that are aligned and then polymerized to form anisotropic films have also been reported for transmission [47] and for reflective mode TN displays [48]. Also biaxial compensation films made from stretched films have been used to give improved viewing angles.

Multidomain twisted nematic

In this device each pixel is divided into either two or four domains by using photoalignment (referred to as linear photo-polymerization–LPP) rather than rubbing techniques (see section C2.3.3) to align the liquid crystal into two or four directions. The observed viewing angle is now an average of the two or four different viewing angles [12, 49]. This gives a more symmetrical and overall wider viewing cone but due to disclinations between these domains the maximum contrast is usually lower.

C2.3.7.2 Reflective mode twisted nematic devices

Devices with two polarizers are commonly used in multiplex TN displays (watches, calculators, etc) but single polarizer devices with an internal reflector can offer better brightness, higher contrast, crisper images (due to reduced parallax that is required for high information content displays) and colour. When driven by an active matrix such devices could be used in low power information devices, hand held computers, digital cameras, games displays, mobile phones etc.

The TN device is a good contender for these displays, as the technology to make them is well known on current production lines. There are competing devices such as active matrix addressed dyed devices [50, 51] and optically compensated birefringence displays [52–54], PDLC [55], holographic PDLC [56] and chiral nematic displays [57].

The twist of the nematic phase can be either low (e.g. ~52°) or high (~190°); within these types there are many options [58]. Most require one or more retardation films to provide good achromatic black and white states. Colour can be introduced by using a birefringence mode and while the colours are bright there is no red colour and no grey levels, thus colour filters are usually used even though light is lost due to absorption in them. The light passes through the colour filters twice thus they are much less absorbing than in transmissive displays; indeed, they are often made even less absorbing so as not to minimize the light absorption. Thus weaker colours are usually found. If a plane aluminium mirror is used as the reflector then a diffusing film that may also give some light direction, to make the device appear brighter, is used. Alternatively, a structured reflector [59] that gives some light redirection and reduces specular reflection is used. In this way, the reflectance within a smaller cone angle can be increased from 11 to 18%. The contrast ratio of a reflective TN is in the region of 5:1 with a reflectance of up to 20%.

C2.3.8 Alternatives to the twisted nematic device

Many display modes are known and just a few of these are described that are either important or illustrate a new feature. With the reduced cost of manufacture of active matrix backplanes, it is now possible to consider alternative liquid crystal modes to the TN mode that were formerly not possible because they could not be multiplex addressed (the voltage/transmission curve was not steep enough) and thus required direct drive.

C2.3.8.1 Vertically aligned mode

Liquid crystals with a negative $\Delta\varepsilon$ find use in this display mode. In an electric field, these compounds align with their short molecular axis parallel to the applied electric field. The molecular reorientation caused by an electric field is shown in figure C2.3.13. The basic cell consists of molecules that are homeotropically aligned (i.e. at approximately 90° to the substrate) which between crossed polarizers gives a very good black state. Upon applying an electric field the molecules tilt and eventually become parallel to the substrate. Removing the voltage gives a reversal of this effect. In practice a small tilt from homeotropic alignment is required, i.e. 89° pretilt so that the molecules tilt in one direction, otherwise a random tilting occurs that gives a speckled appearance between crossed polarizers.

This mode was originally called the 'deformation of aligned phase' (DAP) mode [60] but is now more commonly referred to as a vertically aligned nematic (VAN) display. These devices typically have a short response time and good contrast ratio but the small pretilt to give uniform operation can reduce this. Because there is a practical limit on how negative the $\Delta\varepsilon$ can be, the drive voltage tends to be slightly higher than for TN cells; this can be a drawback for low voltage use.

A development of this device incorporates a compensation film to compensate the retardation created by having a small tilt from homeotropic to regain the good black state [61]. To achieve the wide

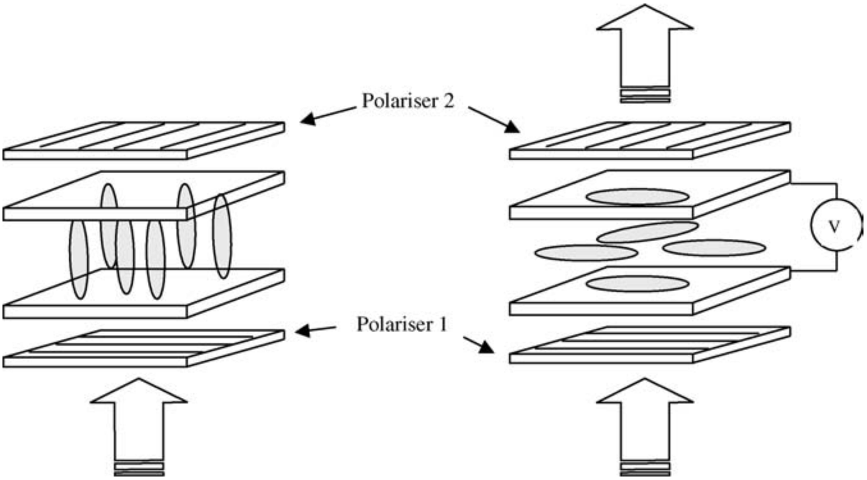


Figure C2.3.13. Schematic drawing of the molecular orientation in the on and off states of a VA mode display.

viewing angle required in monitors etc, small triangular protrusions built from photoresist are formed such that each face gives the molecules on that face a preferred direction of tilt [62]. When an electric field is applied, domains are produced each giving a different viewing cone. The average of these is a symmetrical and wide viewing cone. Such displays have been made into full colour PC monitors and televisions that have become available in 2003.

C2.3.8.2 In plane switching mode

While known for some time this device mode was developed into a display in 1995 [63]. The molecules, which can be of either positive or negative $\Delta\varepsilon$, are aligned with a very small pretilt angle ($<2°$); ITO electrodes are interdigitated on one surface only (figure C2.3.14). Polarizers are positioned such that the first one is in line with the rubbing direction of the electrode plate, and the other is twisted at 90° to the first polarizer; in this state no light is transmitted. When a field is applied to the liquid crystal the molecules in the middle of the cell move towards one of the electrodes; the molecules near the surface are pinned to the surface so do not move easily. This action generates a twist towards the centre of the cell and the liquid crystal acts as a birefringent plate and light is transmitted. Only at very high voltages are all the molecules rotated. Grey levels arise due to varying the voltage and thus the twist.

The advantage of this mode is that the molecules always remain in the plane of the cell and this gives very good viewing angles as at no time is the viewer looking along the long axis of the molecules which is usually the cause of poor viewing characteristics. The problem with this device is that there are two electrodes per pixel together with their transistors, which take up a lot of surface area such that the aperture ratio (a measure of the area that is available for light to be transmitted) is only 40%, creating brightness problems. The response times that were once a problem have now been improved such that they can show video frame rate pictures. These displays have now become commercially available in PC monitors and televisions.

C2.3.8.3 Fréedericksz cells

This was the first type of liquid crystal display mode to be found [64]. The molecules (of positive $\Delta\varepsilon$) are aligned homogeneously (small pretilt angle) and when a field is applied move to align with their

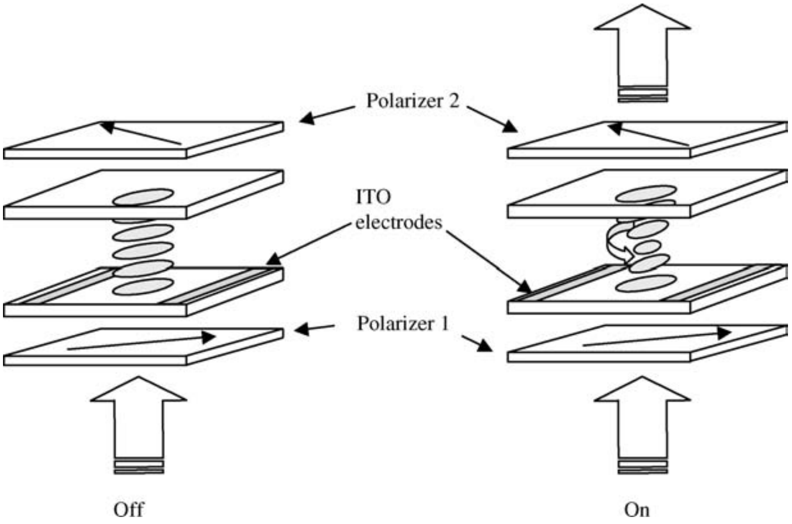


Figure C2.3.14. Schematic drawing of the molecular orientation in the on and off states of an IPS mode display. This shows how the electric field from the ITO electrodes on one surface causes the molecules in the middle of the liquid crystal film to twist.

long axes parallel to the electric field (figure C2.3.15). The molecules at the surfaces do not move and lead to a small residual retardation such that a good black state is not achieved without the use of a compensation foil to correct this residual retardation. They are not used in displays but are used as variable retarders as the retardation of the cell between crossed polarizers is voltage dependent. The retardation at $V = 0\,\mathrm{V}$ is $d\Delta n$ and for the black state between crossed polarizers it is $d\Delta n = \pi/2$ (where d is the cell gap).

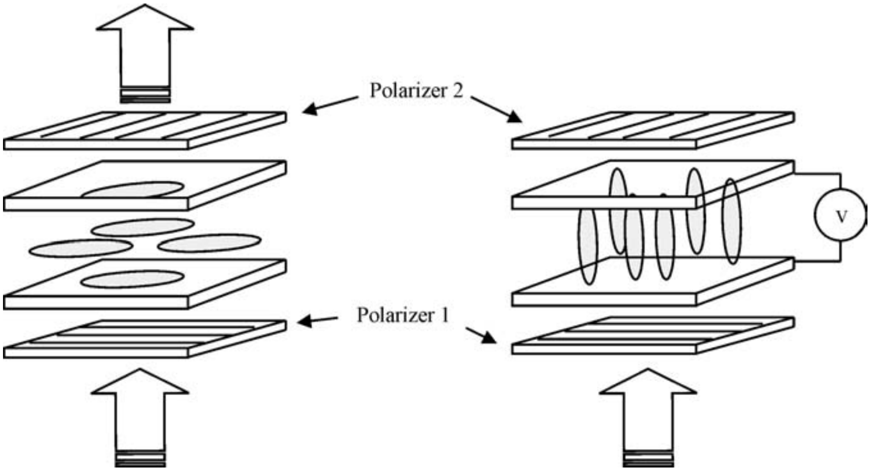


Figure C2.3.15. Schematic drawing of the molecular orientation in the on and off states of a Fréedericksz mode display.

A reflective version, in which a mirror, rather than a polarizer, is placed at a critical distance behind the cell so as to perfectly reflect the correct phase of light, has been described recently [31, 65]. In this way, the reflective device behaves like a transmissive cell in the normally black mode but has half the cell gap.

C2.3.8.4 Pi cells (or π-cells)

These cells are unusual in that, for nematic liquid crystals, they exihibit very fast response times, often 1–2 ms compared to >20 ms of comparable materials in the TN device. Bos and Koehler [66] found that by using a particular orientation of molecules the backflow of the liquid crystal in the switch-off mechanism of a TN cell could be avoided and much shorter response times generated. Figure C2.3.16 compares the switch-off mechanism of these cells and shows a schematic transmission/time curve for this process. It is seen that the flow within the cell is all in the same direction in pi cells and in opposite directions in TN cells. This allows pi cells to switch off faster.

The high pretilt angle alignment of the molecules is not easy to achieve and switching ideally occurs between a partly off state and partly fully on state. The change in optical retardation gives the contrast for which compensation foils between the cell and the polarizers are required to give good black and white operation. Thus a bias voltage rather than zero volts is used. The switching is also found to require a point defect on the surface in each pixel to start the process of switching [67]. Thus while this device has many attractive features, in practice it is difficult to use, but active matrix drive full colour displays with short response times have been demonstrated using this effect which is now often referred to as the optically controlled birefringence (OCB) mode.

C2.3.8.5 HAN cells

The hybrid aligned nematic (HAN) cell is an example of how changing the light path in a display can help with the manufacturing process. This birefringence mode is usually considered for use in reflective

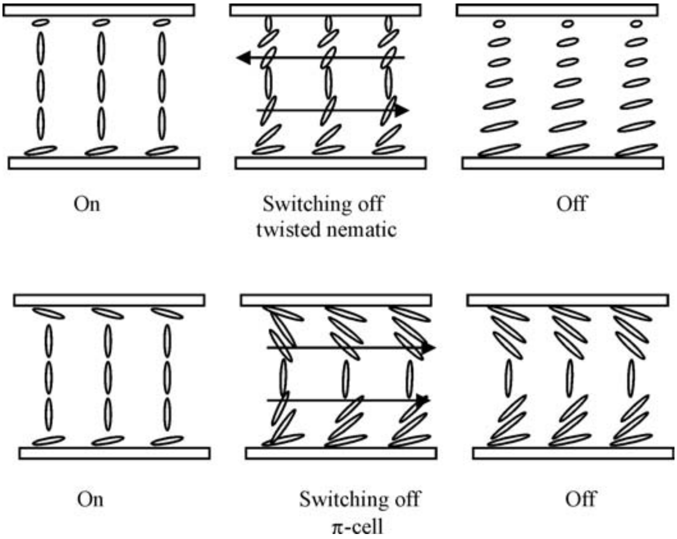


Figure C2.3.16. When the field is removed from a pi cell the molecular flows are in the same direction. It is faster to switch to the off state than an equivalent TN cell whose molecular flows oppose each other.

displays because the retardation (Δnd) it can provide at the mirror surface (in the off state) is equal to $\lambda/4$ thus the cell gap (d) required for black/white contrast is $\lambda/2\Delta n$, twice that of Fréedericksz or VA mode cells; the thicker cell gap allows easier fabrication. One surface of the cell is treated with a low pretilt polyimide and the other has 90° pretilt thus the molecular orientation in the non-driven state looks like half of a pi or OCB cell. Upon applying a voltage, the molecules become homeotropic and between crossed polarizers give a black state. This mode is also called the reflective optically controlled birefringence mode (R-OCB) and has been developed into an amorphous silicon TFT addressed full colour demonstration display [68]. Response times of <12 ms and high brightness are also reported.

C2.3.8.6 Axially symmetric aligned microcell mode

When a photopolymerizable material is dissolved in a thin film of liquid crystal and the film irradiated, through a mask in discrete regions with UV light, a polymer forms and gradually draws the polymerizable material to the irradiated regions. Thus a polymer wall is formed. In this way a small cell, the size of a pixel, can be formed. The alignment of the liquid crystal within this cell is defined by interactions on the cell walls as well as the substrates such that a complex but very reproducible structure is formed. This effect has been made into displays initially using a liquid crystal of positive $\Delta\varepsilon$ [69] and then with negative $\Delta\varepsilon$ materials that gave improved contrast and viewing angle [70]. Large area full colour displays with 300:1 contrast and <30 ms response times having very wide viewing angles have been made and are commercially available.

C2.3.9 Dyed displays

Rather than using polarized light to observe the molecules switching, dichroic dyes can be used. Dichroic dyes often have elongated molecular structures similar to liquid crystal molecules; in one direction they appear coloured and in another direction (ideally at 90°) they appear colourless (usually they are less coloured). When these dyes are dissolved in a liquid crystal their long axes align with the liquid crystal long axes. When the direction of the liquid crystal molecules changes the dye also moves, hopefully such that it also shows a dramatic change in colour. If the dye is longer than the liquid crystal molecules, thermal fluctuations of the smaller liquid crystal molecules are averaged out. However, the dye will still absorb some light even when fully switched; this reduces the transmission of the clear state. The light absorbing state is optimized such that all polarizations of the light become absorbed; low Δn liquid crystals are used to prevent wave-guiding (i.e. the Mauguin limit is exceeded, equation (C2.3.5)). In some cases, displays with two or three layers have been suggested to provide more absorption in all polarization directions [50].

There are many types of dyed display [71, 72]; three of these are representative. The first was suggested by Heilmeier and Zanoni [20] and is made by dissolving a dichroic dye in a positive $\Delta\varepsilon$ nematic liquid crystal in either a Fréedericksz cell or a TN cell with the simple polarizer arranged such that its light is absorbed by the dye and a dark state is formed. On applying a voltage the liquid crystal (and dye guest) become homeotropic and do not absorb as much light; the cell appears brighter. The voltage/transmission curve is very shallow so multiplex addressing is very limited. The twisted mode is usually used; it gives much wider viewing angles than the non-dyed TN and leads to its use in transmissive mode indoor information signs (as in airports etc). Response times are long as the dye increases the liquid crystal viscosity.

The second mode, suggested by White and Taylor [73], also uses a positive $\Delta\varepsilon$ liquid crystal but this time it is a chiral nematic with a pitch length of a few microns, and a polarizer is not required. Almost any aligning layer can be used, low or high pretilt. The twisted structure maximises the light absorption of the off state (figure C.2.3.17 shows the low pretilt version). In the on state the homeotropic state is

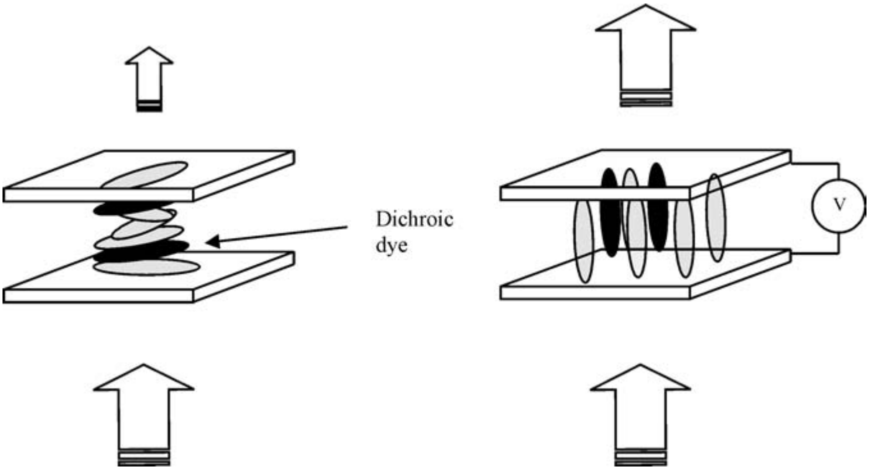


Figure C2.3.17. Schematic drawing of the molecular orientation in the on and off states of a White and Taylor mode dyed display. The off state is dark and the on state light. Polarizers are not required.

produced and thus the cell is almost clear. Due to the tighter helical twist that has to be unwound the drive voltage is often 15 V or more. It is usually used as a reflective device with a contrast ratio of about 25:1 and a reflectance of 20%. In transmission the contrast ratio is 6:1. It finds applications in avionics displays that require bright wide viewing angle displays.

While known for some years the third device mode was used in a recent display that illustrated the use of a dyed negative $\Delta\varepsilon$ chiral nematic liquid crystal [51]. Figure C2.3.18 shows a schematic drawing of this device in which the black dyed long pitch chiral nematic phase is aligned homeotropically (the helix is unwound by the surface forces) and appears clear. When a voltage is applied the molecules move to lie homogeneously and the dye now absorbs light; this is enhanced because the chiral nematic now twists and is able to absorb all polarizations of the incident light. This was used as a reflective device over an

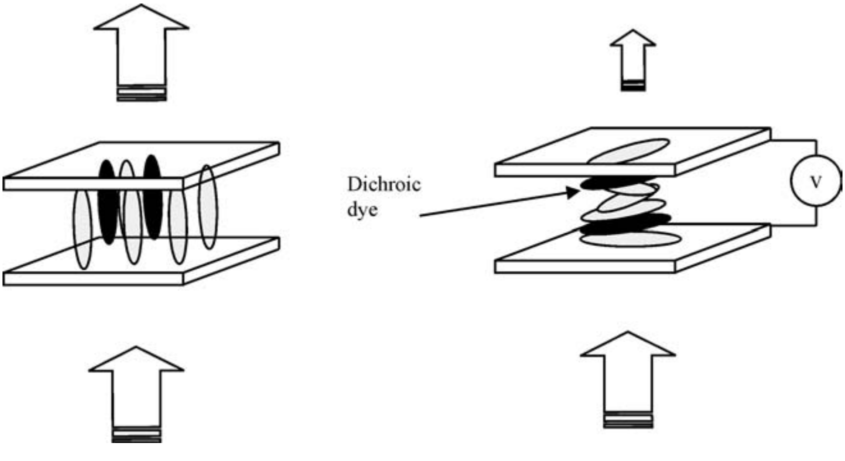


Figure C2.3.18. Schematic drawing of the molecular orientation in the on and off states of a dyed vertically aligned chiral nematic mode display. The off state is bright and the on state dark.

active matrix backplane with colour filters to provide a contrast ratio of 5:1 and 26% reflectance; this compares with $<20\%$ of an optimized single polarizer, reflective colour TN device. However, due to the high viscosity of the dyes it has longer response times than the non-dyed equivalents. Dyes are not often used in active matrix addressed displays as they lower the liquid crystal resistivity, but in this case the device could tolerate this.

C2.3.10 Multistable and bistable devices

Bistable devices offer two major benefits, the first being that once a pixel is switched it will remain switched so that the addressing scheme can move to other pixels. Thus, in principle, an infinite number of rows of pixels can be passively addressed. Very often the pulses of electricity used to switch the pixel are shorter than the optical response time of the liquid crystal such that short row addressing times can be achieved. Additionally, there is no requirement to keep any field on the pixels once they are addressed so lower power consumption can be achieved assuming that the data do not need changing very often. This is often offset against the feature that most bistable devices require higher drive voltages than monostable displays so more energy is used to change the image.

C2.3.10.1 Chiral nematic displays

Chiral nematic liquid crystals with a short helical pitch length can reflect 50% of a specific waveband of light (equations (C2.3.1) and (C2.3.2)); the other 50% is transmitted along with all other wavelengths—a black background can be used to absorb the transmitted light so that the observer only sees the reflected colour. This reflection only occurs when the helix is well aligned in what is known as the planar texture. When these helices are jumbled up (the focal conic state) they do not reflect light but give a weakly scattering state that looks transparent and with a black background appears black. Using an electric field on the molecules of positive $\Delta\varepsilon$ it is possible to switch between these two states via the homeotropic state (all molecules parallel to the applied field) each of which is stable [74] (figure C2.3.19). The stability of the focal conic state in particular can be enhanced by optimizing the surface forces (surface stabilized chiral nematic) or a polymer network can be made within the liquid crystal (polymer stabilized chiral nematic) [75]. There are also many stable intermediate states that give as many as 22 grey levels, thus the device is multistable. By stacking three cells on top of each other [57, 76], each reflecting a particular colour (red, green or blue), it is possible to reflect either white light, when all reflect in the planar texture, or each colour can be selected to give that particular colour or give black when all the cells are in the focal conic state. In this device, all the pixels can be any colour so it is particularly effective at giving brighter colours than conventional reflective displays using colour filters where a red picture for example can use only one third of the pixels of that area. In the white state a reflectance of about 40% is achieved with a black/white contrast of $>10:1$.

These displays have found use in display boards usually as single colour (green on black) displays but are now made into large area full colour advertising boards as large as 6 m × 3 m [77].

C2.3.10.2 Bistable nematic displays

These displays aim to provide passively addressed displays with high duty ratios while still retaining the good contrast and viewing angle of direct drive TN displays. They would have applications in portable displays such as mobile phones, PDAs and information tools. As low power is the attractive feature they should also work well in reflection. Several variants have been developed and have different ways to generate the bistability [78].

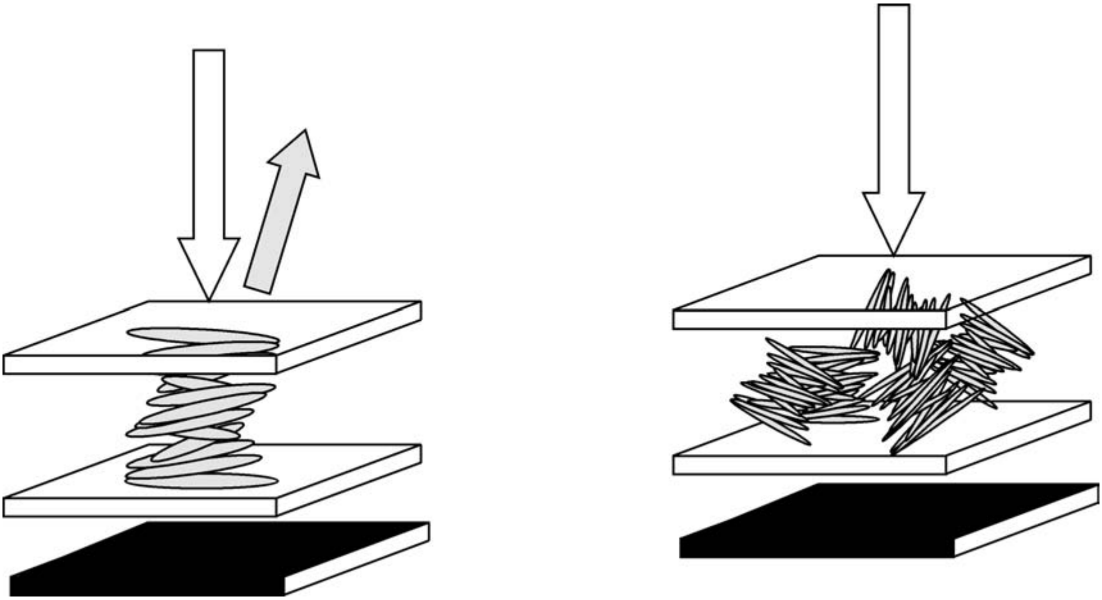


Figure C2.3.19. A reflective surface stabilized chiral nematic mode display exhibits two extreme states. In the planar state (left) the helix is oriented such that it can reflect a waveband of visible light while the rest is transmitted and absorbed by a black absorber. The device appears coloured. In the focal conic state (right) the helices are small and of random orientation; most light is now transmitted and then absorbed by the black absorber.

A zenithal bistable device uses a grating structure on one surface and homeotropic alignment on the other. The liquid crystal can have either a twisted [79, 80] or non-twisted structure [81] between the electrodes. Figure C2.3.20 shows the basic device effect that switches between a homeotropic and a splayed state. A more recent twisted mode using a nematic liquid crystal of negative $\Delta\varepsilon$ can be reconfigured by an electric field polarity reversal. An applied field couples with the flexoelectricity of the splayed state to allow switching. It is therefore an example that uses the flexoelectric properties of the liquid crystal. These stable states are formed by addressing with very short ($<100\,\mu$s) pulses of <20 V

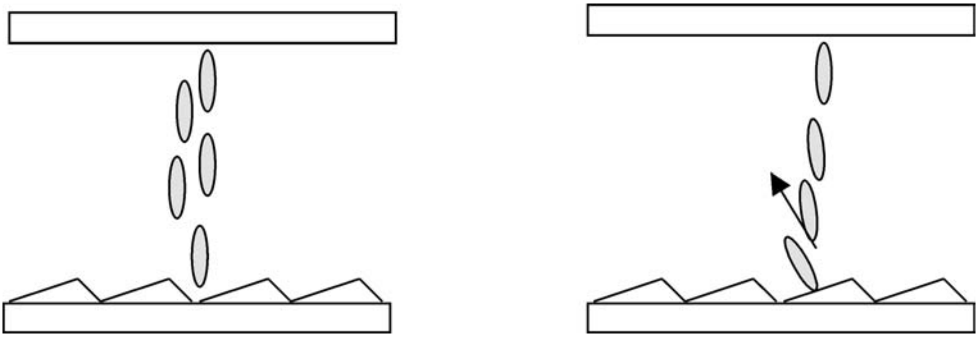


Figure C2.3.20. Schematic drawing of the molecular orientation in the two stable states of a zenithal bistable device showing the grating structure that helps stabilize the states. The arrow indicates the flexoelectric dipole direction of the splayed orientation.

or longer pulses having a low voltage (3–8 V) can be used. The device is neither too dependent on cell gap nor has it the need for compensation films to achieve good viewing angles. It has been made in both transmissive and reflective mode using two polarizers. The two states are stable even when the cell is physically pressed.

The second variant relies on one alignment surface having a much lower surface energy than the other such that at a modest voltage the alignment of the liquid crystal molecules on the weaker surface can be broken and the alignment from the other surface propagated through the cell [82, 83]. Initially, this was achieved by using evaporated SiO_2 on one surface, but polymer films are being developed for mass production use. Like the zenithal device it is also switched using short pulses of about 20 V but can also be driven with long voltage pulses. It can be made in transmissive or reflective variants. The reflective variants are clearly more important as the low power device is better used without a backlight to give an overall low power display module.

The viewing angle and contrast of these displays rivals STN displays and both are candidates for low power, reflective displays. A possible problem with reflective mode devices of both these displays is that small cell gaps ($<3\,\mu$m) are currently required—these are commercially difficult to achieve in high yield.

A further bistable mode variant uses a 360° TN liquid crystal similar to STN displays in which the 0, 180 and 360° twist states are all stable. Between crossed polarizers the 0° state is black and the 360° state white, but the 180° state is the lowest energy state. By building polymer walls around the pixel the 180° state is prevented from forming. These states can be realized by electrical addressing [84].

None of these displays is commercially available yet but they are very likely to be so in future.

C2.3.11 Light scattering displays

The first commercial liquid crystal display was a dynamic scattering device [85] in which the negative $\Delta\varepsilon$ liquid crystal was doped with ions that under a low frequency electric field would oscillate between the ITO electrodes. This motion disturbed the long liquid crystal molecules such that they formed micron size domains that created light scattering. Because the liquid crystal was of negative $\Delta\varepsilon$ it was held parallel to the ITO plates while the ion movement tried to disrupt it. It was called 'dynamic' because the scattering becomes less intense when the field is removed. Using a high frequency field, which the ions cannot respond to, produces a clear state. It is no longer used. Scattering displays have been reviewed [86].

Of more importance is the family of polymer dispersed liquid crystal devices suggested by Fergason [87] and popular in the 1990s; several reviews [88, 89] and books [90] have been published. Typically, these devices switch between a scattering unpowered state and a powered clear state. However, devices of the opposite mode have also been made using small amounts of a polymer network formed *in situ* within the liquid crystal film [55, 91]. Figure C2.3.21 shows a typical PDLC device in which micron sized droplets of a liquid crystal (usually nematic but it can be chiral nematic or smectic) are encapsulated in a polymer matrix. The director of the liquid crystal is random from droplet to droplet and thus a refractive index mismatch occurs at the boundaries that causes light scattering. The most common way to make these films is to UV cure a mixture of acrlyated monomers, oligomers and liquid crystal; the acrylated monomers and oligomers polymerize and the liquid crystal is forced out of solution and forms droplets. The other major technique is to coat an aqueous emulsion of latex, liquid crystal and surfactants and then evaporate off the water to give a thin light scattering film.

When a field is applied to the film the nematic phase director in the droplet orientates to lie parallel with the field; no longer does a refractive index mismatch occur between droplets so the film appears transparent. These films require about 30–40 V for a 20 μm thick film. One advantage of this device is that when about 50% of the material is polymer it can be coated onto plastic substrates and large areas made on a roll to roll coater. This led to its use as a window film giving privacy and diffuse lighting.

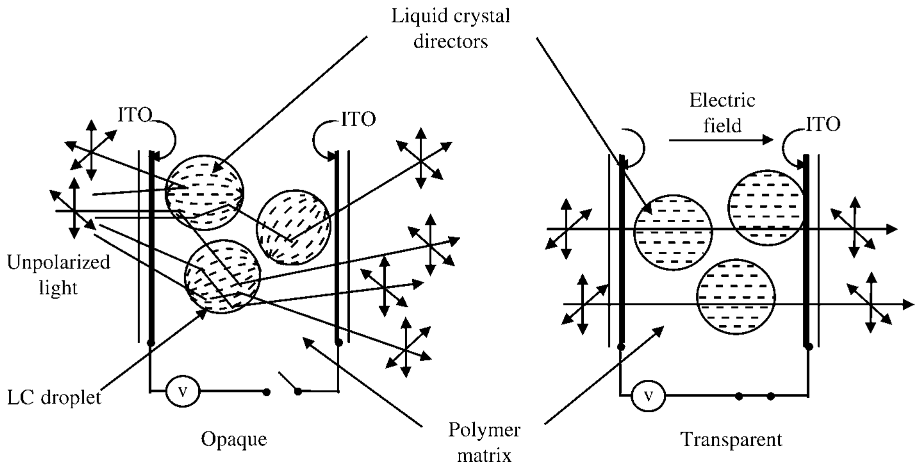


Figure C2.3.21. Droplets of liquid crystal within a polymer matrix cause light scattering; when an electric field is applied to the film the liquid crystal molecules' long axes align with the field and the film does not scatter light.

When 80% liquid crystal is used the film is made between ITO coated glass plates; this is a low voltage film intended for use in projection and direct view displays. These films operate at about 6–8 V and are usually about 6 μm thick. An active matrix can drive these low voltage films; many prototypes were made but perhaps the most advanced of these was a novel internal reflection device giving full colour but only 8% reflectance [92]. PDLC films in general cannot be addressed using multiplexing techniques because of the very shallow voltage/transmission curves.

At present, due to the high cost of the large area films relative to what customers want to pay, PDLC is only used in a few niche applications and in displays. They have failed to make a breakthrough.

In recent years, holographically formed scattering layers have been made that can selectively reflect light and may form the basis of a display [56] or it may find use in optical telecommunication devices [93].

C2.3.12 Smectic based liquid crystal displays

C2.3.12.1 Non-chiral smectic phases

In the early 1980s the search for new liquid crystal effects moved to smectic liquid crystals; many effects were found and were usually bistable because the liquid crystal was too viscous to flow back after the voltage was removed as it does in nematic displays. Very often the device had to be heated to cause reversal of the display. However, one electrically reversible scattering to clear mode [94] was made into a high resolution monitor display [95] but it was overtaken by the advent of active matrix addressed displays. Also of particular note is the work of Kahn *et al* [96] to develop laser addressed displays that also gave a contrast between scattering and clear from which rewritable photomasks were made. These smectic displays and their uses have been reviewed [97, 98].

C2.3.12.2 Chiral smectic phases

The chiral smectic C phase was predicted and demonstrated to be ferroelectric [99]. While the bulk phase is very viscous a very low viscosity moderates the path that the molecules take when electrically

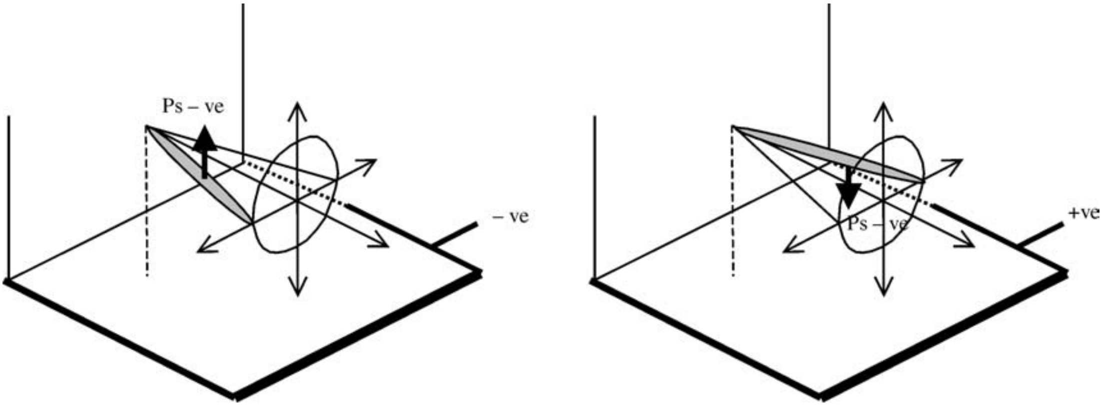


Figure C2.3.22. Ferroelectric liquid crystal device showing the spontaneous dipole moment (P_s) of the molecule having a charge that is either attracted or repelled by the charge applied to the electrodes. To change position the molecules move within the smectic layers and they describe a cone. Ideally the cone angle is 45°.

switched. Figure C2.3.22 shows how one molecule, which is part of a layer structure (as in Figure C2.3.1) rotates around a cone when switched by an electric field. Note that the molecule has a large permanent dipole moment (shown by the arrow) that can interact strongly with an electric field leading to short response times. After being electrically poled the molecules are constrained to align and move co-operatively with their neighbours. In the ideal case the molecules switch in a plane through 45° and when viewed between crossed polarizers give maximum contrast of black to white (absorption to transmission of light). It can also be bistable or can be held in a position by a small retaining voltage separate to the addressing voltage, thus many pixels can be addressed and in principle it is an infinitely multiplexable passive drive display. This discovery has stimulated many years of research [6] to define the best ways to use this effect and design molecules that exhibit both a smectic C phase with a large dipole moment (called a spontaneous dipole) and low viscosity over a temperature range that allows its use in displays.

These displays exhibit very short response times (e.g. 50–100 μs rather than 20–70 ms for nematic displays; this allows moving images to be shown). Due to this very short response time it also allows sequential colour in which a backlight flashes sequentially red, green and blue light and to coincide with these illuminations the display shows the image for red, green and blue sequentially. As these occur many times each second the eye averages them. Thus, no colour filters are required and each pixel is any colour. It also operates at a low voltage ($\geqslant 5$ V) and because it is an in plane switching device it has a wide viewing cone. In principle, these features should have ensured its success. But the chiral smectic C phase is very difficult to align and retain in a well-aligned state. There are many possible orientations each having different properties and in some modes can be unstable to mechanical shock. Electrical addressing is also complex. Some mono colour and 15 inch high resolution colour displays were developed [100] and offered on the Japanese market only. While of excellent appearance they lacked the robustness of AMTN displays. Later using a different electrical drive scheme, Itoh *et al* [59] showed a 17 inch full colour panel operating at 30 V with 12 μs per line addressing but its contrast ratio (60:1) was lower than required. The cell gap was 1.3 μm, that made its manufacture very difficult.

However, with the process improvements and lower cost of silicon it is becoming cost effective to use silicon itself instead of glass as one substrate and this is used to incorporate the complex circuitry to drive the FLC device. Due to the high cost of silicon only very small displays are viable (less than 1 inch diagonal) but they still contain > 0.5 million tiny pixels (XGA and SXGA resolution). These can be used in digital camera displays, ‘near to the eye’ viewfinders and projection devices to form very large (30–50 inch) images on a screen [101, 102].

However, even with all the positive attributes the problems in aligning and driving FLC displays remain and it has not achieved more than entry to niche markets.

C2.3.13 Projection displays

Liquid crystal displays can be used in both front and rear projection applications. It is usually cheaper to project a large (>40 inch) image than make a direct view display, hence the interest in projection displays. The problem of achieving high contrast together with high brightness is a critical issue. Most liquid crystal display projectors use a transmissive active matrix but three such cells are required. Competing against this is the LCOS technology in which a small reflective microdisplay (<1 inch) is used; indeed usually three such displays are used, one for each colour channel although the aim is to make a liquid crystal device that is fast enough such that only one display will be required using sequential colour illumination and addressing. FLC can achieve this but the light level is lower than for nematic devices due to addressing features of FLC displays. Recent developments in electrical addressing and projection techniques also allow colour sequential nematic liquid crystal displays to use one LCOS device [103]. For front projection they compete against an impressive deflecting mirror technology that offers brighter, higher contrast displays but at higher cost. Thus liquid crystal displays could fill this lower cost market which means that even smaller (lower cost) silicon based microdisplays may be seen in the future. In rear projection liquid crystal displays are more commonly used, as brightness is less critical.

C2.3.14 Future prospects

Liquid crystal displays are now the ultimate for flat screen displays; the scope of applications they can address is extremely broad. Their reign started by creating markets that the then best technology (CRTs) could not address and gradually started to replace the CRT with better products in the applications where it had the most problems. Now the liquid crystal display competes head on in the CRT’s main market segment (monitors and TVs). Yet still after 30 years of liquid crystal display technology the CRT remains overall the major display mode because in most cases where the two compete it is much cheaper, but it cannot get cheaper while liquid crystal displays probably can.

There are now competing flat panel technologies (such as organic light emitting diodes—OLEDs) that offer fast addressing times and are not as limited by temperature as liquid crystal displays. Will these replace liquid crystal displays in the near future? Bearing in mind the CRT story this seems unlikely even without technical arguments but simply due to the massive infrastructure and investment in liquid crystal displays. It is more probable that OLEDs will start by having success in areas where liquid crystal displays are not that good such as in wide temperature range applications. There is debate about which of the two uses less power for portable applications. However, the final choice will be that of the consumer and which display gives the best value. Most large liquid crystal display makers are also developing OLED displays as they did with liquid crystal displays when they only made CRTs. It is possible that new faster light emitting displays can open markets that liquid crystal displays cannot address so easily, thus we may see new applications for flat panel displays.

Looking at the liquid crystal displays themselves there are more exciting new modes being taken to full colour demonstration than ever before. These attempt to overcome the speed issues and have ever-lower power consumption using bistable low voltage modes in reflection. Thus we shall see improved monitors and television screens; the much quoted wall mounted television is here but most houses are not designed for it; they are designed for CRT televisions. However, flat screen monitors are being used and liked in the desk environment; it cannot be long before flat screen television liquid crystal displays replace portable or second TVs. In large size televisions they will also compete with plasma screens. Projection liquid crystal displays operate in a tough market segment but many believe they can survive. Plastic displays sound attractive but when considered seriously no one has a killer application that must have plastic rather than glass; until this occurs the market will be small and thus the plastic devices will be expensive.

References

[1] Reinitzer F 1888 *Montasch Chem.* **9** 421

[2] Toyne K J 1990 *Thermotropic Liquid Crystals*, *Critical Reports in Applied Chemistry* vol 22, ed G W Gray, Chapter 2

[3] Toyne K 1998 *Handbook of Liquid Crystals* vol 2A, ed D Demus (New York: Wiley–VCH) p 47

[4] Coates D 1992 *Liquid Crystals Applications and Uses* vol 1, ed B Bahadur (Singapore: World Scientific) p 91

[5] Sage I C 1992 *Liquid Crystals Applications and Uses* vol 3, ed B Bahadur (Singapore: World Scientific) p 301

[6] Lagerwall S 1998 *Handbook of Liquid Crystals* vol 2B, ed D Demus (New York: Wiley–VCH) Chapter VI-2

[7] Hartshorne N H 1974 *Microscopy of Liquid Crystals* Monographs in Microscope Series vol 48 (London: Microscope)

[8] Hirschmann H and Reiffenrath V 1998 *Handbook of Liquid Crystals* vol 2A, ed D Demus (New York: Wiley–VCH) Chapter 3.1

[9] DeJue W H 1980 *Physical Properties of Liquid Crystalline Materials, Liquid Crystal Monographs 1* (London: Gordon and Breach)

[10] Schadt M and Muller F 1978 *Inst. Elec. Electron. Eng., Trans. Electron. Devices* **25** 1125

[11] Le Comber P G *et al* 1979 *Electron. Lett.* **15** 179

[12] Schadt M 1996 *Nature* 381

[13] Chaudhari P, Lacey J A *et al* 1998 *Japan. J. Appl. Phys.* **37** L55

[14] Yamasaki T, Kawakami H and Hori H 1996 *Colour TFT Liquid Crystal Displays*, ed SEMI Standard FPD Technology Group

[15] Ouderkirk A, Weber M E, Benson U, Cohn S, Jonza J M, Wortman D L and Stover C A 1997 *US Patent* 6 543 153

[16] Wortman D L 1997 *SID International Symposium Digest of Technical Papers* M-98

[17] Broer D, Lub J and Mol G N 1995 *Nature* **378** 467

[18] Coates D, Goulding M J, Greenfield S, Hanmer J M W, Marden S A and Parri D 1996 *SID International Symposium Digest of Technical Papers (Applications)* p 67

[19] Kelker H and Scheurle B 1969 *J. Phys Collique* **C4** 104

[20] Heilmeyer G H and Zanoni L H 1968 *Appl. Phys. Lett.* **13** 91

[21] Schadt M and Helfrich W 1971 *Appl. Phys. Lett.* **18** 127

[22] Gray G W, Harrison K J and Nash J A 1973 *Electron Lett.* **9** 118

[23] Hulme D S, Raynes E P and Harrison K J 1974 *J. Chem. Soc. Chem Commun.* 98

[24] Coates D, Gray G W and McDonnell D G 1978 *US Patent* 4 113 647

[25] Eidenschink R, Erdmann S, Krause J and Pohl L 1977 *Angew. Chem.* **89** 103

[26] Boller A, Cereghetti M, Schadt M and Scherrer H 1977 *Mol. Cryst. Liq. Cryst.* **42** 1225

[27] Gooch C H and Tarry H A 1974 *Electron Lett.* **10** 2

[28] Pohl L, Weber G, Eidenschink R, Baur G and Fehrenbach W 1981 *Appl. Phys. Lett.* **38** 497

[29] Alt P and Pleshko P 1974 *IEEE Trans. Electron. Devices* **21** 146

[30] Bradshaw M J and Raynes E P 1983 *Mol. Cryst. Liq. Cryst.* **99** 107

[31] Leuder E 2001 *Liquid Crystal Displays* (Chichester: Wiley)

[32] Yeh P and Gu C 1999 *Optics of Liquid Crystal Displays* (Chichester: Wiley)

[33] Scheffer T J, Nehring J, Kaufmann M, Amstutz H, Heimgarten D and Eglin P 1985 *SID International Symposium Digest of Technical Papers* p 120

[34] Waters C M, Raynes E P and Brimmel V 1985 *Mol. Cryst. Liq. Cryst.* **123** 303

[35] Pauluth D and Wachtler A E F 1997 *Chirality in Industry II*, ed A N Collins (New York: Wiley) p 264

[36] Okamura O *et al* 1987 *Japan. J. Appl. Phys.* **26** L1784

[37] Odai H *et al* 1988 *International Display Research Conference Digest of Technical Papers* p 195

[38] Schadt M, Buchecker R, Leenhouts F, Boller A, Villegar A and Petrzilka M 1986 *Mol. Cryst. Liq. Cryst.* **139** 1

[39] Scheffer T J and Clifton B 1992 *Proc. SID* **23** 228

[40] Ozeki M, Mori H and Shidoli E 1996 *SID International Symposium Digest of Technical Papers* p 107

[41] Fujita S, Yamaguchi H, Mizuno H, Ohtani T, Sekime T, Hatanaka T and Ogawa T G 1999 *J. SID* **7/2** 135

[42] Brody T P *et al* 1974 *SID International Symposium Digest of Technical Papers* p 166

[43] Kaneko E 1998 *Handbook of Liquid Crystals* vol 2A, ed D Demus (New York: Wiley–VCH) Chapter 3.2

[44] Ono N, Ushiki T, Suzuki K, Ushiyama T, Noda Y, Kamikawa T, Kaneko K and Morozumi S 1990 *SID International Symposium Digest of Technical Papers* p 518

[45] Tarumi K, Bartmann E, Geelhaar T, Schuler B, Ichinose H and Numata H 1995 *Asia Display* p 1

[46] Mori H, Itoh Y, Nishiura Y, Nakamura T and Shinagawa Y 1997 *SID International Symposium Digest of Technical Papers* p 941

[47] Van der Witte P 1999 *Japan. J. Appl. Phys.* **38**

[48] Uesaka T, Toyooka T and Kobari Y 1999 *SID International Symposium Digest of Technical Papers* p 95

[49] Schadt M 1999 *Proc. Eurodisplay* p 27

[50] Hasegawa, Takeda K, Sakaguchi Y, Tarra Y, Egelhaaf J, Leuder E and Lowe A C 1999 *SID International Symposium Digest of Technical Papers* p 962

[51] Kretz T, Gomez G, Lebrun H, Coates D and Reaney S, 2002 *SID International Symposium Digest of Technical Papers* p 798

[52] Uchida T, Nakayama T, Mujashita T, Suzuki M and Ishinabe T 1995 *SID International Symposium Digest of Technical Papers* p 599

[53] Uchida T, Ishinabe T and Suzuki M 1996 *SID International Symposium Digest of Technical Papers* p 618

[54] Uchida T, Ishinabe T and Miyashita 1998 *SID International Symposium Digest of Technical Papers* p 774

[55] Fujisawa T, Nakata H and Aizawa M 1996 *SID International Displays Research Conference* p 401

[56] Kato K, Tanaka K, Tsuru S and Suki S 1994 *J. SID* **2/1** 37

[57] Davis D, Kahn A, Huang X Y, Doane J W and Jones C 1998 *SID International Symposium Digest of Technical Papers* p 901

[58] Tillin M D, Towler M J, Saynor K A and Beynon E J 1998 *SID International Symposium Digest of Technical Papers* p 311

[59] Itoh Y, Fujiwasa S, Kimura N, Mizushima S, Funada F and Hijikigawa M 1998 *SID International Symposium Digest of Technical Papers* p 221

[60] Shiekel M F and Fahrenschon K 1971 *Appl. Phys. Lett.* **19** 393

[61] Yamauchi S, Aizawa M and Clerc J F 1989 *SID International Symposium Digest of Technical Papers* p 378

[62] Takeda A and Kataoka S *et al* 1998 *SID International Symposium Digest of Technical Papers* p 1077

[63] Oh-e M, Ohta M, Aratani S and Kondo K 1995 *Asia Display* p 577

[64] Fréedericksz V and Zolina V 1933 *Trans. Faraday Soc.* **29** 1978

[65] Leuder E, Muecke M and Polach S 1998 *Asia Display* p 173

[66] Bos P J and Koehler K R 1984 *Mol. Cryst. Liq. Cryst.* **113** 329

[67] Koma N, Miyashita T and Uchida T 1999 *SID International Symposium Digest of Technical Papers* p 28

[68] Shibazaki, Ishinabe T, Miyai M, Yoshida K, Ugai Y, Miyashita T and Uchida T 1998 *Asia Display* p 51

[69] Yamada N, Kohzaki S, Funada F and Awane K 1995 *SID International Symposium Digest of Technical Papers* p 575

[70] Kume Y, Yamada N, Kosaki S, Kisishita H, Funadu F and Hijikigawa M 1998 *SID International Symposium Digest of Technical Papers* p 1089

[71] Bahadur B 1992 *Liquid Crystals Applications and Uses* vol 3, ed B Bahadur (Singapore: World Scientific) p 68

[72] Bahadur B 1998 *Handbook of Liquid Crystals* vol 2A, ed D Demus (New York: Wiley–VCH) p 257

[73] White D L and Taylor G N 1974 *J. Appl. Phys.* **45** 4718

[74] Greubel W, Wolff U and Krüger H 1973 *Mol. Cryst. Liq. Cryst.* **24** 103

[75] Yang D K and Doane J W 1992 *SID International Symposium Digest of Technical Papers* p 759

[76] Hashimoto K, Okada M, Nishiguchi K, Masazumi N, Yamakawa E and Taniguchi T 1998 *SID International Symposium Digest of Technical Papers* p 897

[77] Magink Display Technologies 2003 *Marketing* 4 April

[78] Dozov I 2003 *SID International Symposium Digest of Technical Papers* p 946

[79] Jones J C, Wood E, Bryan-Brown G and Hui V C 1998 *SID International Symposium Digest of Technical Papers* p 858

[80] Wood E and Bryan Brown G P 2000 *SID International Symposium Digest of Technical Papers* p 124

[81] Bryan-Brown G 1997 *SID International Symposium Digest of Technical Papers* p 37

[82] Joubert C, Angele J, Boissier A, Davi P, Dozov I, Elbhar T, Pecout B, Stoenescu D and Vercelletto R 2002 *SID International Symposium Digest of Technical Papers* p 30

[83] Martinot-Lagarde Ph, Dozov I, Polossat E, Giocondo M, Lelidis I, Durand G, Angele J, Pecout B and Boissier A 1997 *SID International Symposium Digest of Technical Papers* p 41

[84] Bos P, Watson P, Anderson J E, Sergan V and Hoke C D 1999 *Proc. Eurodisplay* p 397

[85] Heilmeyer G H, Zanoni L H and Barton L H 1968 *Appl. Phys. Lett.* **13** 46

[86] Bahadur B 1992 *Liquid Crystals Applications and Uses* vol 1, Chapter 3, ed B Bahadur (Singapore: World Scientific)

[87] Fergason J L 1984 *US Patent* 4 435 047

[88] Coates D 1995 *J. Mater. Chem.* **5** 2063

[89] Doane W 1992 *Liquid Crystals Applications and Uses* vol 1, ed B Bahadur (Singapore: World Scientific) p 362

[90] Drzaic P 1995 *Liquid Crystal Dispersions* (Singapore: World Scientific)

[91] Hikmet R 1991 *Liquid Crystals* **9** 405

[92] Sonehara T, Yazaki M, Iisaka H, Tsuchiya Y, Sakata H, Amako J and Takeuchi T 1997 *SID International Symposium Digest of Technical Papers* p 1023

[93] Bunning T J, Natarajan L V, Sutherland R L and Tondiglia V P 2000 *SID International Symposium Digest of Technical Papers* p 121

[94] Coates D, Crossland W A, Morrissey J H and Needham B 1978 *J. Phys. D Appl. Phys.* **11** 2025

[95] Crossland W A and Cantor S 1985 *Proc. SID* **16** 124

[96] Kahn F, Kendrick P N, Leff J, Livioni J, Loveks B E and Stepner D 1987 *SID Digest of Technical Papers* **18** 254

[97] Coates D 1992 *Liquid Crystals Applications and Uses* vol 1, ed B Bahadur (Singapore: World Scientific) p 275

[98] Coates D 1998 *Handbook of Liquid Crystals* vol 2A, ed D Demus (New York: Wiley–VCH) p 470

[99] Meyer R J, Lieber L, Strzelecki L and Keller P 1975 *J. Physique Lett.* **3** L9

[100] Mizutani H *et al* 1997 *International Conference on Ferrolectric Liquid Crystals* p 66

[101] Akimoto O and Hashimoto S 2000 *SID International Symposium Digest of Technical Papers* p 194

[102] Birch M, Kruerke D, Yates C, Macartney A, Peden D and Coates D 2002 *SID International Symposium Digest of Technical Papers* p 954

[103] Brennesholtz M S 2002 *SID International Symposium Digest of Technical Papers* p 1346

C2.4

Technology and applications of spatial light modulators

Uzi Efron

C2.4.1 Introduction

Spatial light modulator (SLM) technology has made tremendous progress in the past few years. Thus, device technologies such as active matrix-driven liquid crystal devices (LCDs) and ferroelectric LC (FLC) devices as well as micro-mirror SLMs, regarded as exotic novelties less than two decades ago, are now commercially available and constitute a significant share of the display market.

Regarded as 'dream-devices' only a decade ago, the multiple quantum well (MQW) SLMs are now a reality with array sizes of 256 × 256 and staggering frame rates of over 100 kHz.

It is interesting to note that the trend in applications continues to be dominated by the display applications. Optical interconnects (OIs), an almost unknown application at the time, have moved strongly forward in the last few years with the explosive development of communications, while the exotic optical data processing application is still awaiting the opportunity for a break-through.

This chapter, aimed at describing the main SLM technologies and applications is organized as follows.

The scope and definition of the main SLM types is detailed in section C2.4.2. The bulk part of this review contained in section C2.4.3 describes the main types of SLM currently available or those in an advanced development stage. Particular emphasis is placed on the recent emerging technologies including LC, micro-mirror or micro-electro-mechanical-systems- (MEMS-) based arrays as well as MQW devices. Short summaries of the 'older' SLM technologies, including solid state electro-optic, magneto-optic as well as acousto-optic and photorefractive devices are also presented in section C2.4.3. Novel SLM concepts including electro-holograms and photonic band-gap (PBG) materials are presented in section C2.4.4. Section C2.4.5 presents the main SLM applications including displays, optical communication, optical data processing, programmable diffractive optical elements, adaptive optics and wavelength image converters. Finally, section C2.4.6 presents some of the fundamental limits as well as the future trends of this technology.

There have been numerous reviews in the form of books, book-chapters and conference proceedings covering SLM technology, since the mid-1970s. Some of the more recent reviews (since 1990) are given in [1–11].

C2.4.2 SLM definition and general description

SLMs are devices that spatially modulate the amplitude, phase or the polarization state of an optical beam. By 'spatial modulation', we mean either a one-dimensional (1D; linear array) or a two-dimensional (2D) spatial modulation. Although not yet practically implemented, a three-dimensional (3D) spatial modulation of an optical beam (e.g. a dynamic 3D hologram) is also possible.

A combination of the modulation parameters such as phase–amplitude or polarization–amplitude modulators has also been demonstrated. Since the most important modulation parameter is the amplitude or intensity of the optical beam, the modulation of either the phase or the polarization state of the beam is often used to affect amplitude (intensity) modulation through the conversion of the spatial phase or polarization modulation into intensity modulation using interferometric arrangement or crossed polarizer–analyser configurations, respectively. The control or addressing of the SLM is another important parameter. The main choices here are electrical addressing (e.g. electro-optical or electro-absorption SLM) and optical addressing (e.g. optically-addressed SLM or OASLM); since electronic addressing is the most practical choice, this form is sometimes used to excite an intermediary field that in turn affects the optical property of the beam. Examples for this intermediary addressing are: (electro)-magneto-optic SLM (MO-SLM) and (electro)-acousto-optic SLM. The next important parameter is the readout beam configuration. We normally distinguish between transmission-mode and reflective-mode devices. Next, we specify the readout operational wavelength of the device. Thus, in addition to the commonly used visible regime, SLM devices exist and operate in the IR and UV regimes of the electro-magnetic spectrum.

Figure C2.4.1 shows a cross-section of a liquid crystal SLM. According to the categories defined above, the device is a reflective-mode, photo-activated (PA) SLM, based on a crystalline Si photo-substrate (in a Schottky-diode configuration) and a nematic LC layer as the electro-optical modulator. The device accepts visible–IR input imagery (400–1100 nm) and can modulate the readout beam to form dynamic imagery from UV to long IR, with frame rates of up to 1 kHz. This particular device has been operated in various LC configurations allowing both phase and amplitude modulations of the readout beam.

Table C2.4.1 summarizes the main characterization parameters of SLM devices. The main physical mechanisms utilized in SLMs are listed in table C2.4.2.

Table C2.4.3 summarizes the main specification and performance parameters of SLMs.

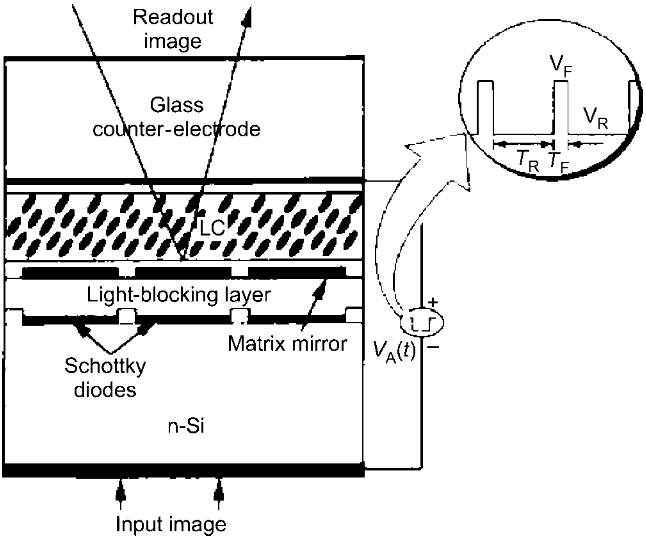


Figure C2.4.1. A reflective-mode, Si-Schottky-diode, LC SLM. (After [12].)

Table C2.4.1. Main characterization parameters of SLMs.

SLM parameter	Parameter range	Example
Spatial dimension	1D, 2D, 3D	
Modulation type	Amplitude, phase, polarization conversion	
Addressing mode	Electrical, optical (PA); (electro)-acousto-optic, (electro-)magneto-optic	
SLM configuration	Transmissive, reflective	
Operational wavelength	Visible, IR, UV, (mm-wave)	
Optical modulator type	Liquid crystal, MEMS, MQW, photo-refractive, acousto-optic, magneto-optic	LCD, Bragg cell
Driver type	Active matrix (Si), passive matrix (Si), CCD (Si, GaAs), photo-conductor (Si, GaAs, BSO, . . .)	Si-backplane

Table C2.4.2. Physical mechanisms employed in SLMs.

Physical mechanism	Examples
Electro-optic (linear)	PLZT [13], (KD*P) [14, 15], MSLM (LiNbO3) [16, 17], ferro-electric liquid crystal (FLC) [18]
Electro-optic (quadratic)	Nematic-LCD [19], cholesteric liquid crystals [20]
Electro-mechanical (MEMS)	Deformable mirror (DMD-TI) [21, 22], grating LV [23, 24], deformable membrane devices [25–27]
Electro-absorption	GaAs (CCD-SLM) [28, 29], (MQW-SLM) [30]
Electrostatic phase distort	Thin oil film LV [31], gel-SLM [32]
Electro-capillarity	Bistable mercury-mirror devices [33, 34]
Photo-refractive	PROM/PRIZ/PICOC [35–40], photo-MQW-SLM [41, 42]
(Electro)-acousto-optic	Bragg cells (1D) [43], acousto-MQW [44]
(Electro)-magneto-optic	Magneto-optic (MO-SLM) [45]
Thermal phase changes	VO_2 modulator (IR) [46], LC-smectic/cholesteric transition [47], bubble cross-connect switch [48]
Plasmon-electro-optic enhancement	[49]
PBG modulation	[50]
Active light emission	OLED-array [51, 52], VCSEL array [53, 54]

C2.4.3 Main SLM types and their performances

C2.4.3.1 Liquid crystal devices

LC-SLMs: scope

As pointed out in the introduction, the main application of LCDs and, for that matter, of most SLM devices in general, is still that of displays (see section C2.4.5 later). LCs are particularly suitable for display applications featuring a combination of extremely high electro-optic coefficients, broad-band visible operation and response times, which will match human vision. Although developed mostly

Table C2.4.3. Main SLM specifications and performance parameters.

Performance parameter	Description	Units
Array size	No. of horizontal/vertical pixels	N/A
Resolution		LP/mm @ MTF
Temporal response	Rise/decay time, frame rate	ms, Hz
Spectral bandwidth	Input/output beam	nm
Input sensitivity (PA Device)	Min. photo-activating power	$\mathrm{cm}^2\,\mathrm{mW}^{-1}$, $\mathrm{cm}^2\,\mathrm{J}^{-1}$
Contrast ratio/phase dynamic range		dB; bits; radians
Diffraction/output modulation efficiency		%
Output uniformity		%
Physical array size		cm × cm
Power consumption		W
Operating temperature range		°C
Relative humidity range		%
Shock resistance		G

for displays, LC-SLMs were and are still being studied for other applications as well. These include optical data processing, OIs, adaptive optics and programmable Kinoform (binary-optical) elements, such as variable focal length lenses and beam-steering devices (see later in section C2.4.5, SLM-applications).

Historically, the nematic materials were the first to be used for SLMs and other display applications (such as watch displays). As described earlier, a nematic LC layer operating as the electro-optic modulator is used in either a twisted-nematic (TN) [55] or in a controlled-birefringence configuration [56]. The main advantage of the TN configuration is in its optically broad-band operation and the relatively lower sensitivity to the LC thickness variations.

Nematic LCDs are mainly used as either small (1–3 inches) panels (or 'light valves') for projection displays or as larger (10–15 inches) direct view panels for personal digital assistants (PDAs), lap-top and hand-held PCs (HPCs), or mobile phone displays. The LCDs for both applications are manufactured in both transmissive and reflective configurations. Earlier versions of LC projection displays were based on PA configurations, consisting of a photo-conductive layer used to convert the incoming, low-light-level input image (e.g. from a small-size CRT) into a spatial voltage distribution across the nematic LC layer. A separate beam is then used to read out the LC modulation and project the image onto a large screen. Due to the rapid progress in large CMOS array technology, most projection LCDs of today are based on electrically driven or electrically addressed (EA) LCDs. The older generation, PA LCDs, also called 'liquid crystal light valves' (LCLVs) [57], are still in use for some of the high-end projection display applications requiring a combination of high resolution and high brightness such as very large screen displays and electronic cinema.

Electrically addressed LCDs were first based on the use of a passive X–Y addressing scheme [58]. See section C2.3 by Coates. Although this concept relies on the threshold voltage property of LC materials, the MULTIPLEXING level (e.g. the number of pixels in a line) is quite limited. With the advance in VLSI technology allowing large transistor driver arrays to be manufactured at high yield, these passively addressed LCDs are now replaced by actively addressed LCDs (based on active transistors in each pixel) in most applications.

Two other important developments in LCD technology took part in the 1980s and 1990s. These are: the discovery and development of FLC [59] and the development of a scattering mode, polymer-dispersed LC (PDLC) [60, 61]. The FLC, as the name suggests, is based on a LC material class (smectic C*) having a permanent electrical dipole moment and thus the molecules can physically be flipped with the reversal of the electric field. This is in contrast to the nematic LC, which responds only to the magnitude of the field (E) and not to its polarity, with subsequent longer response time. The PDLC is based on LC 'bubbles' or 'droplets' immersed in a polymer matrix such that upon the application of an electric field, their average director orientation changes from a random distribution to a fully-field-aligned one. By properly matching the refractive index of the LC to that of the polymer, the result is the formation of a scattering mode in the off-state and a transparent mode, with application of the field. The most important aspects of this new LC form are (1) the *polarization-independent* optical modulation capability and (2) the simple packaging allowing pre-processed PDLC sheets to be pre-manufactured and assembled into the LCD without the need for the cumbersome vacuum-filling of cells with liquid crystal material.

The last decade (1990–2000) has witnessed additional important developments in LCD technology among which are the development of an efficient poly-silicon-based driver circuitry [62], as well as the in-plane switching mode (IPS) [63] and more recently a novel photo-polymerization alignment technique [64]. The poly-silicon driver technology allows an efficient electronic driver to be fabricated on the top of a glass-based substrate, thus enabling the fabrication of high performance, large panel LCDs for lap-top computers. The IPS mode significantly increases the viewing angle for an LCD display. Finally, the novel contact-less, photo-alignment (PAL) method allows the formation of multi-LC domains to reduce polarization dependence.

The principal LC types and modes of operation

As the name implies, LCs represent an intermediate phase combining the physical properties of both solids and liquids. Specifically, this class of materials behave as anisotropic liquids, as they have no translational or long range lattice order and in this regard they are fluids or liquids. However, both their electrical and optical properties are directionally dependent or anisotropic. It is important to keep in mind that the LC phase is an intermediate phase, which exists only within a narrow temperature range (a few tens of degrees). The LC material solidifies below this range of temperature, while undergoing a liquid–liquid transition above this temperature range, where it becomes a normal isotropic liquid. The LC anisotropy is tied to the directionality in the (average) physical orientation of the molecules.

The degree of anisotropy can be described by introducing the *LC director* and its orientation L, which indicates the average, macroscopic orientation of the LC molecules.

There exist three main types of LC class, namely, nematic, cholesteric and smectic mixtures. A brief description of each class and its mode of operation is given in the following. For an extensive review of LC classes and their associated physical and electro-optic properties, see references [65–70].

Nematic liquid crystals

In this class (or phase) the aggregate of the LC molecules show directional anisotropy in both their dielectric and optical properties. We thus have longitudinal and transverse polarizabilities $\alpha_{\|}$ and $\alpha_{\perp}$, dielectric constants $\varepsilon_{\|}$ and $\varepsilon_{\perp}$, as well as refractive indices $n_{\|}$ and $n_{\perp}$. All these parameters are defined with respect to the long molecular axis, which practically coincides with the average orientation of the LC director. Positive-dielectric anisotropy NLCs are characterized by having: $\Delta\varepsilon = \varepsilon_{\|} - \varepsilon_{\perp} > 0$, and similarly for $\Delta\alpha$ and Δn.

It should be noted that both positive and negative dielectric and refractive index anisotropies can be found in the LCs. Of particular interest are NLCs having a negative $\Delta\varepsilon$ but positive Δn.

These are sometimes referred to as: 'negative-anisotropy NLCs'.

The director orientation coincides with the optical axis of the refractive index ellipsoid. Thus, one usually substitutes the ordinary (n_o) and the extraordinary (n_e) refractive indices for $n_\perp$ and $n_{||}$, respectively.

NLCs do not have a permanent dipole moment. Thus, the application of an electric field $\boldsymbol{E}$ beyond a threshold level E_c induces a dipole moment μ, proportional to $\boldsymbol{E}$. The field then exerts a torque proportional to: $|\mu \times \boldsymbol{E}| \sim |\boldsymbol{E}|^2$, which re-orients the LC molecules or, macroscopically, the LC director L. This re-orientation results in an effective tilting of the refractive index ellipsoid, thus varying the effective birefringence, as experienced by the incident polarized optical beam.

The threshold field, E_c, is given by: $E_c = (\pi/d)\surd[k/\varepsilon_0\Delta\varepsilon]$. Here, d is the thickness of the LC cell and k is the appropriate elastic constant. One can define a thickness-independent *threshold (or critical) voltage*, $V_{\text{th}} = E_c d$ (rather than a threshold field), beyond which the applied voltage starts re-orienting the LC director. The turn-on and turn-off response times, τ_{ON} and τ_{OFF}, are given by: $\tau_{\text{ON}} \propto \tau^* d^2/[(V/V_{\text{th}})^2 - 1]$, where: $\tau^* = \gamma/\pi^2 \boldsymbol{k}\lambda\tau_{\text{OFF}}$.

The quadratic dependence of the torque exerted on the LC director by the applied voltage results in two important characteristics of NLCs:

(1) The NLC's orientation responds to the *magnitude* or RMS value of the applied voltage and is therefore insensitive to the polarity of the applied voltage or field.

(2) The turn-on response time is inversely proportional to V^2, whereas the off-state or the turn-off time is independent of the applied voltage. This sets the lower limit for the total response time of NLCs in practical devices to milliseconds. Such response is ideal for display devices, but is too slow for optical processing, as well as most telecommunication switch (cross-connects) applications.

There exist three basic configurations of NLC cells:

(1) Planar configuration in which the NLC molecules are homogeneously aligned parallel to both electrode planes in a certain *alignment* direction, which can be physically imposed by a particular treatment of the LC cell electrodes.

(2) Homeotropic configuration in which the NLC molecules are uniformly aligned *perpendicular* to the cell electrodes.

(3) A twist configuration in which the parallel alignment orientations of the NLC molecules at the entrance and exist cell electrodes or windows differ by an angle ϑ_T.

This configuration is formed by fabricating alignment layers with different orientations on the two cell electrodes. The NLC molecules then align in planes parallel to the cell electrodes, where in each plane the director orientation is rotated (or twisted) relative to the preceding plane. This results in the director re-orientating or twisting continuously from the alignment orientation at the entrance electrode, to the alignment direction at the exit electrode. Typical twist angles are 45° (hybrid-field effect); 90°, 180° (π-cell) and 270° or above (super-twisted NLC = STN configurations) [71].

LC alignment

As is obvious from the previous discussion, one must define the LC alignment within the cell in order to make use of its electro-optical properties. This is accomplished by generating alignment 'marks' at the boundary cell surfaces, which then 'anchor' the adjacent LC molecules in those pre-defined directions, thereby affecting the alignment of the bulk LC material within the cell. Obviously, such alignment via anchoring to the surface cannot be sustained over large distances. Thus, practical LCDs are physically constrained to a thickness of a few tens of micrometres (seldom exceeding 20 μm).

There exist several methods of LC alignment. The most important ones are:

(1) Shallow angle (SiO_x) deposition (SAD) [72].

(2) Mechanical rubbing or buffing (BUF) [73].

(3) Photo-alignment (PAL) [74, 75].

(4) Langmuir–Blodgett (LBL) process [76].

(5) Ion-beam etching (IBE) of polymer films [77].

(6) Formation of a grating relief structure (GRS) [78].

The relative advantages/disadvantages of each are given in table C2.4.4.

Parallel or homogeneous alignment. This is the simplest form of alignment in which the surfaces of both electrodes are aligned in the same direction. This alignment usually employs positive-dielectric anisotropy LC material, whose director (normally parallel to the surface and in the direction of the alignment marks) would tend to tilt towards the electrodes, upon the application of an electric field or voltage.

In order to avoid the formation of opposite tilt-angle domains, a small (1–2°) pre-tilt angle with respect to the electrode surface is usually defined to be at a 180° difference between the two electrodes. Since in this alignment the LC molecules are anchored parallel to the electrode surfaces, the alignment

Table C2.4.4. Summary of LC alignment methods.

Alignment method	Advantages	Disadvantages
SAD	Excellent tilt control	Complex process, difficult to mass-produce
BUF	Simple, demonstrated production	Requires mechanical contact, difficult tilt control
PAL	Non-contact, amenable to mass-production, remotely programmed orientation, multi-domain	Low anchoring energy, tilt control, long-term stability
LBL	Natural method for homeotropic alignment	Difficult to mass-produce, complex process, tilt control
IBE	Good tilt control, demonstrated production	
GRS	Lithographic process	Complex process, difficult to mass-produce, optical losses (scattering)

mechanism makes use of the largest foot-print of the elongated LC molecules and is therefore the most powerful (highest anchoring energy).

Homeotropic LC alignment. The homeotropic or tilted-perpendicular alignment (TPA) of LCs is a particularly difficult alignment, both due to the small foot-print of the molecules on the aligning surface, and the difficulty of a precise control of the pre-tilt angle (around 88–89°), which is crucial for a proper operation of LCDs[†].

This form of alignment, as well as some of its derivatives (such as the hybrid alignment), are attractive for LCD display technology, as they can lead to a very high contrast and reduced angular sensitivity. Several methods of implementing the homeotropic alignment are given in [79–88].

The electro-optic effect in NLC cells

The nematic LC molecular orientation becomes distorted under the application of an applied voltage (the so-called Fredericksz transition), with the LC director tilting at an angle $\vartheta(V)$. Although the tilt-angle varies across the LC cell thickness, one can approximate the variation of the effective birefringence $\Delta n_{\mathrm{eff}}(V)$ of the cell by:

$$\Delta n_{\mathrm{eff}}[\theta(V)] = n_{\mathrm{e}} n_{\mathrm{o}} / \{ n_{\mathrm{o}}^2 \cos^2[\theta(V)] + n_{\mathrm{e}}^2 \sin^2[\theta(V)] \}^{1/2} - n_{\mathrm{o}}$$

assuming a spatially uniform tilt-angle, $\vartheta(V)$, throughout the cell.

In the case of a *parallel-aligned configuration*, assuming a positive dielectric anisotropy LC, the director will be re-oriented from a parallel alignment ($\theta(V) = 0°$, for: $V < V_{\mathrm{th}}$) to an almost perpendicular alignment with respect to the cell electrodes ($\theta(V) = 180°$), at sufficiently high field or voltages (typically: $V > 10\ \mathrm{V_{RMS}}$). This change which results in $\Delta n_{\mathrm{eff}}[\theta(V)]$, changing from: $\Delta n_{\mathrm{eff}} = \Delta n_{\max} = n_{\mathrm{e}} - n_0$ to: $\Delta n_{\mathrm{eff}} = 0$, can be used to effectively rotate the polarization plane of the incoming optical beam. Thus, the intensity of an incoming beam *polarized at 45°* to the alignment direction of the (parallel-aligned) LC will be proportional to: $I = I_0 \sin^2(\Delta\varphi/2)$, following a crossed-analyser behind the LC cell. The retardation angle, $\Delta\varphi$, can be approximated by: $\Delta\varphi = 2\pi d\{\Delta n_{\mathrm{eff}}[\theta(V)] - n_{\mathrm{o}}\}/\lambda$, using the uniformly aligned LC cell approximation, with a tilt-angle given by $\theta(V)$.

By aligning the input polarization *along with* the LC alignment direction, this configuration can be used for a *pure phase modulation*, where the voltage-modulated phase, $\Delta\varphi$, is given by the same expression as above.

The same basic applications for intensity (polarization-rotation) modulation or a pure phase modulation can be accomplished using the homeotropically aligned NLC arrangement discussed earlier. In this case, one needs to use a negative-anisotropy LC in order to accomplish a director re-orientation by the application of an electric field.

The effect of continuously controlling the effective birefringence in a homogeneously aligned cell (whether parallel or, perpendicularly aligned) is termed the ‘controlled birefringence’ or CB-effect.

The TN configuration is formed by adding a chiral (helically structured) chemical agent into a nematic mixture. In this configuration, the orientation of the polarization plane of an optical beam, polarized either parallel or perpendicular to the alignment at the entrance window, will follow the director orientation rotation caused by the twisted structure across the LC cell, exiting with the plane of polarization aligned with the LC alignment direction at that exit window.

[†] For homogeneously aligned (or planar-aligned) LC configuration, the pre-tilt angle guards against the formation of multiple domains in the LC layers, which is cosmetically unacceptable for displays. However, for the perpendicular TPA alignment, the pre-tilt control is much more critical, in that it determines the final in-plane orientation of the LC director. This orientation must be predetermined and controlled in order for a display device to function properly.

This optical rotation effect, which occurs in the absence of an electric field, is subject to the condition: $U = \pi d\Delta n/\lambda\theta_T \gg 1$ (limit), where θ_T is the total twist angle in the cell.

Defining the rotatory power, P_R, as the fraction of the optical beam intensity whose polarization gets rotated along with the twist angle as previously described, it can be shown that this rotatory power can be approximated by: $P_R \approx U^2/(1 + U^2)$, for both regimes of: $U \gg 1$ and $U \ll 1$ [68, p 372, 89, 90]. In particular, for $U \gg 1$, the rotatory power level is almost unity, showing that the polarization plane of the incoming beam is almost perfectly rotated with the twist. The effect of applying an electric field is to tilt the director perpendicular to the (almost plane-parallel) layers, towards the direction normal to the cell electrodes. This results in the twist arrangement becoming increasingly distorted and eventually breaking up at sufficiently high fields. The destruction of the twist structure at high field destroys the polarization property and thus allows the use of the locally applied voltage to modulate the polarization rotation or the intensity with the use of a proper polarizer–analyser arrangement.

It should be pointed out, however, that in this arrangement one cannot simply use the effect for pure phase modulation except for very low applied fields [91].

Another important difference between the controlled birefringence and the TN effect is that the latter can be used (subject to the above condition) to rotate the plane of polarization in a wide spectral window (e.g. throughout the entire visible spectrum of 450–650 nm). The CB-effect, on the other hand, is by its very nature a *wavelength-dependent* effect as can be seen by the λ-dependence of the retardation $\Delta\varphi$ on the previous page.

Cholesteric LC

A cholesteric LC (CLC) mixture is formed by the introduction of a relatively large concentration of a chiral agent, resulting in a typical enlargement (numerous 2π cycles) of the twist angle. In the CLC configurations, the degree of the rotational twist is characterized by the period of the twist rotation, namely, the pitch P_0, measured in micrometres, and related to the twist angle θ_T by: $\theta_T = 2\pi d/P_0$. It has been shown [92, 93] that:

1. The pitch, P_0, of the helix formed is inversely proportional to the concentration of the chiral component in the nematic mixture. Typical values of the pitch are between 0.5 and 50 μm and can be attained by varying the chiral component concentration between 20 and 0.5%.
2. The transmission of the planar cholesteric configuration (where the helical axis is perpendicular to the cell surfaces) is high, except in the spectral regions where: $\lambda \approx P_0$, and where a high reflectivity of a circularly polarized light is observed.
3. The twist or helix structure, of the cholesteric phase can be *completely suppressed* by a sufficiently large electric field. In this case, the LC director (for a positive-anisotropy nematic, $\Delta\varepsilon > 0$) lines up with the field (perpendicular to the cell electrodes), with an effective refractive index of n_o, relative to the incoming optical beam.

Upon application of a low electric field, the planar structure turns into a group of randomly oriented helical clusters that strongly scatter light. This scattering phase is termed 'focal conic' texture. The capability of forming both a planar texture (at zero field), which with sufficiently long pitch can be used as a transparent state, and the use of the field-excited focal conic is the basis for the recent interest in using cholesteric materials as a bistable optical modulator. In order to obtain stable, zero-field states for both focal-conic and planar textures, polymer stabilization of the LC structure is used [94, 95].

Smectic materials

Smectic materials possess an ordered *layered* structure. While there exist several subclasses of smectic phases, the two important ones are smectic A and C (SmA, SmC). These differ by the angle in which the molecules tilt with respect to the normal layers. A crystalline order, namely, a well-defined layer period, exists in both the smectic phases. It should be pointed out that the smectic and nematic phases may be based on the same material, of which SmA and SmC may constitute different phases, at different temperature ranges. The degree of disorder in the phases, is increased with increasing temperature, SmC being at the lowest temperature end and, hence, the most ordered phase; SmA follows as a more disordered smectic phase, ending at the highest temperature range with the nematic as the lowest ordered, anisotropic LC phase.

As for the optical properties, SmA is a uniaxial material with its extraordinary refractive index perpendicular to the layers' planes, while the refractive indices in both in-plane axes are equal to the ordinary index. This symmetry, however, is lost in the SmC phase due to the molecular tilt in the smectic planes. Of particular importance is the chiral SmC or SmC* which is formed by adding a chiral (helically structured) chemical agent to the SmC mixture. In this case, the mirror symmetry of the SmC is lost and the formation of a spontaneous dipole moment is possible without the presence of an external electric field. The existence of a permanent dipole moment μ (which is reversible by the application of an electric field) results in a directional (polarity) dependence of the LC director on the externally applied electric field, through the generated torque $\boldsymbol{T}$, where: $\boldsymbol{T} = \mu \times \boldsymbol{E}$. This then leads to on- and off-response times both being inversely proportional to the field: $\tau(\boldsymbol{E}) = \gamma/\mu|\boldsymbol{E}|$. This case, in contrast to the nematic LC behaviour, where the off-time is independent of the applied field, having an effectively lower (field-induced) dipole moment, leads to significantly faster response times of ~10–100 μs.

The first concept and demonstration of a usable FLC configuration was published in 1980 by Clark and Lagerwall [96]. In this configuration, the researchers constructed a thin FLC cell in which the helical structure was suppressed, resulting in a uniform distribution of the director across the LC cell. This arrangement, termed 'surface stabilized FLC' (SSFLC), allows two field-dependent, LC director orientations separated by $2\theta_L$ where θ_L is the tilt- or cone-angle of the FLC. The LC director can switch between both states by the application or reversal of the externally applied electric field. The response times for switching are of the order of 100 μs for typical applied fields (5–10 V μm^{-1}). Despite being limited to binary operation, the SSFLC configuration has been the principal mode of operation for FLC-SLMs for a variety of applications, in particular for LCDs. In this last application [97], pulse width modulation (PWM) methods making use of the fast response of the FLC relative to the required frame time of ~20 ms are used to enable the required grey scale operation.

Another mode of operation is based on using the SmA very near to its SmC transition point. In this configuration (*electro-clinic* effect) [98], the LC director orientation, which is still uniform throughout the cell, tilts proportionately to the applied electric thereby allowing continuous grey scale operation. At the same time, one can attain response times comparable to those of the FLC devices since the LC molecules already possess a permanent dipole moment.

A mode of operation employing SmC* or FLC materials was developed by Funfschilling and Schadt [99] in Switzerland, in 1989. This mode, termed 'deformable helix FLC' (DHF), does not constrain the natural helix formation in a SmC cell, but makes use of the fact that the helix will tend to get distorted upon the application of an electric field. Thus, in the short-pitched helix cell formed (such that: $p \ll \lambda$), this field-induced distortion results in a change in the orientation of the LC director, averaged over the wavelength size LC section observed. This orientation change can be used, similar to the controlled-birefringence effect in a nematic LC cell, for polarization rotation operation.

Very fast speed of response (a few microseconds), as well as grey scale operation, can be attained in this DHF mode, at the expense of a relatively low dynamic range. The latter is a consequence of the averaging effect over the LC director.

Polymer-based LC configurations

Recent advances in both polymerization and lithographic techniques resulted in the introduction of polymers into LC structures. In general, the advantage of introducing polymers into LC mixtures is the formation of mechanically self-supported structures, as opposed to regular LC fluids requiring external mechanical support in the form of cell window glasses or plastics. The first concept, introduced over a decade ago, was that of a PDLC [100–104]. This concept is based on the phase separation which occurs between the LC and polymer materials and which results in the formation of micrometre or, wavelength-size LC droplets within the surrounding polymer matrix. The original concept was based on the use of nematic LC materials. However, in later PDLC versions other LC phases such as FLCs and CLCs are also utilized. In the NLC-PDLC version the electro-optic modulation effect is based on matching the ordinary NLC droplet index of refraction to that of the surrounding polymer matrix. Since, in the absence of an external electric field, the LC director orientation in the droplets is randomly distributed, the droplet index will differ (on average) from that of the polymer matrix. Thus, the PDLC will strongly scatter in this state. Upon application of an external electric field, the LC directors in the NLC droplets become oriented with the field (e.g. parallel to the field direction, for a positive-anisotropy LC). In this state the droplets now appear to have an ordinary index matching to that of the surrounding polymer matrix, for optical beam incident perpendicular to the cell window. This arrangement therefore allows a two-state, polarization-independent, transparent/scattering optical modulation.

Other types of polymer-stabilized LC structure, in particular, a reflective, polymer-stabilized cholesteric texture (PSCT) for a scattering/narrow-band reflective, dual-mode bistable LC modulator, have recently been intensively studied for low-power-consumption displays [105]. One potential application is that of electronic paper.

Liquid crystal SLMs

In the following, we will present a brief description of the main LC-SLM structures. For an extensive review of this technology, the reader is referred to the latest symposia of the Society for Information Displays over the last few years [106] where numerous papers on these structures were published.

Electrically addressed LCDs

A schematic of a reflective-mode, electrically addressed, LCD is shown in figure C2.4.2.

A polarized readout beam enters the device from the top and is reflected from one of the electrode/mirrors placed at the bottom of the device. These electrodes addressed with different voltages affect the LC whose director tilts as shown in proportion to the increasing voltage (from left to right). The alignment layers control the orientation of the LC director at the electrode surfaces.

The LC driving is accomplished using mostly ‘active matrix’ schemes which are based on the use of thin-film circuitry with one or more MOS-based transistors in each pixel node. The circuitry typically employs a row/column drive timing scheme in which each row is sequentially addressed such that within each of the addressed rows the column pixel transistors are sequentially addressed. The MOS transistors in each pixel allow the sequential gating of the signals into each of the pixel’s storage capacitors. The addressing signal, which must be alternating in its polarity (AC) to avoid LC deterioration, is typically 1–10 V RMS.

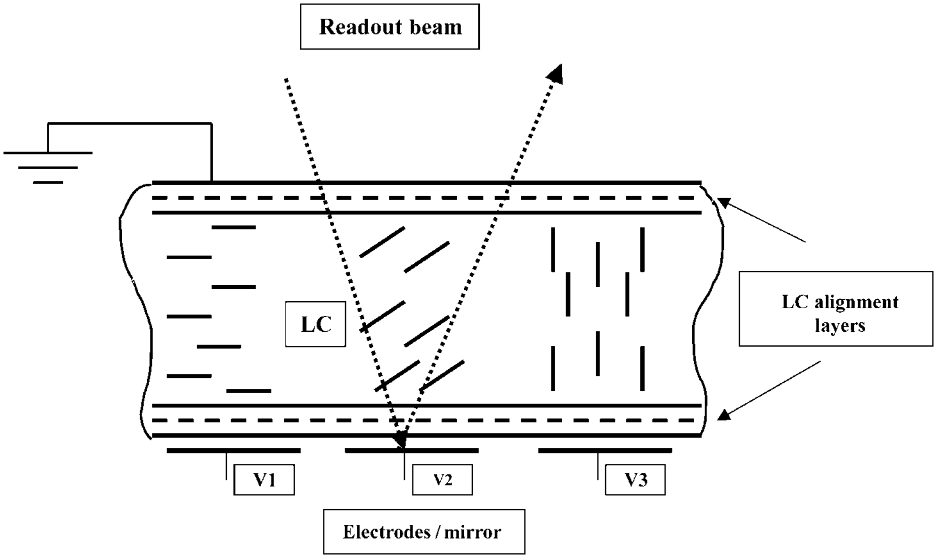


Figure C2.4.2. Schematic of a electrically addressed, reflective-mode LCD.

The silicon substrate in the early direct view LCD panels was amorphous silicon which, due to its limited performance (carrier mobility) has recently been replaced by poly-crystalline Si, grown on a high-temperature glass substrate.

An important consideration for a Si-based driving circuitry is that it must be well shielded from the readout beam. In a transmissive-mode device this implies a mandatory loss in the area- or aperture-factor of the pixels, and therefore in the optical efficiency, as the circuitry region of the pixel cannot be used for light transmission.

This results in an increasingly high loss of optical efficiency for smaller pixel-size arrays, as the (optically shielded) circuitry occupies an increasingly larger portion of the total pixel area. Reflective-mode LCDs, on the other hand, can be made optically efficient regardless of their pixel size.

The LC employed in the direct view displays is typically in a TN configuration, which allows a cell-thickness-independent, broad-band (450–650 nm) operation. Pixel-level-RGB filters deposited on the front glass allow colour operation to be accomplished in conjunction with spatially aligned activation of the RGB sub-pixel drivers. Finally, a crossed analyser at the exit plane, in conjunction with the input polarizer, converts the polarization-rotation at the pixel-level into spatially modulated intensity (with full colour).

As pointed out above, there is a significant structural difference between direct view LCDs (e.g. a lap-top or note-book screen) and projection panel LCDs. While lap-top displays are large-aperture (25–40 cm), transmissive type, projection display panels are typically of smaller size (2–8 cm in linear dimension), usually of a reflective type. These relatively smaller size panels are currently based on a single crystal silicon substrate (the so-called LC on silicon technology (LCOS)). The driver electrodes are usually utilized as the reflection layer with an appropriate metallic coating. The LC mode can either be a hybrid field effect mode [107] first used in the early LCLVs, vertically aligned nematic exhibiting a relatively wide field of view (FOV), or more recently, in-plane-switching configuration [63]. FLC as well as PDLC devices were also demonstrated in small LCD panels for projection displays. FLCs are also currently produced for video camera direct view applications [108]. An ultra-thin-crystalline Si-substrate

Table C2.4.5. Typical performance of electrically addressed LCD projection panel.

Performance parameter	Description	Units
Input/output	Electrically addressed, output intensity modulation	N/A
Device configuration	Reflective mode	N/A
Array size	Up to 2000 × 2000	Pixels
Optical modulator	Nematic liquid crystal (TN configuration)	
Driver array	Single-crystal CMOS array	
Resolution	25	LP mm^{-1} @ 20% MTF
Temporal response	Rise/decay time: 0.5–10 ms, frame-rate: 30–300 Hz	ms, Hz
Spectral region	Visible, 450–650	nm
Input sensitivity (PA device)	N/A	N/A
Contrast ratio	100:1 (40 dB)	dB
Output modulation efficiency	10–30 (unpolarized, collimated beam)	%
Output uniformity	1–3	%
Aperture size	2 × 2 (for a 1024 × 1024 array @ 20 μm pixel-size)	cm × cm
Power consumption	~500	mW
Operating temperature range	10–40	°C
Shock resistance	1–5	G

technology recently developed [109] allows very small transmissive panels to be used for projection and head-mounted display applications.

The above class of electrically addressed LCDs encompasses the use of the different LC configurations discussed earlier: in particular, TN cells, controlled-birefringence operation, as well as FLC [110] and PDLC [111] configurations. Finally, one should mention that in addition to the traditional use of MOS- or CMOS- based active matrix driving schemes, the use of a CCD-based circuitry matrix to address an LCD has also been demonstrated with an array size of 256 × 256 [112].

Typical performance of an electrically addressed LCD panel for projection displays is shown in table C2.4.5.

Photo-activated LCD SLMs

As pointed out earlier, the first type of LC-SLM developed was actually the PA SLM configuration developed at Hughes Aircraft Research Laboratories during the late 1970s [57]. These devices were the first LC-SLM products for both commercial and military uses. A typical reflective-mode, PA SLM structure is shown in figure C2.4.3. The structure of such an SLM commonly known as an LCLV consists of four layers: a photoconductor (PC); a light blocking layer (LBL); a mirror (typically dielectric) (DM) and the LC layer. This structure is sandwiched between two glass electrodes providing both the mechanical support and the necessary sheet-conductivity or surface-electrode function. The operation of the device is as follows: under low illumination level at the input (PC port) most of the bias voltage drops on the (high-resistivity) PC layer, and as a result the LC is biased below threshold level and is consequently not activated. Upon illumination of a region or spot on the photoconductor, this section of the PC becomes conductive and the voltage now shifts to the LC and the LBL/DM layers. The latter, being thin, takes only a relatively small fraction of the AC voltage, which now drops mostly on the LC layer activating those areas aligned with the illuminated sections of the photoconductor. The activation

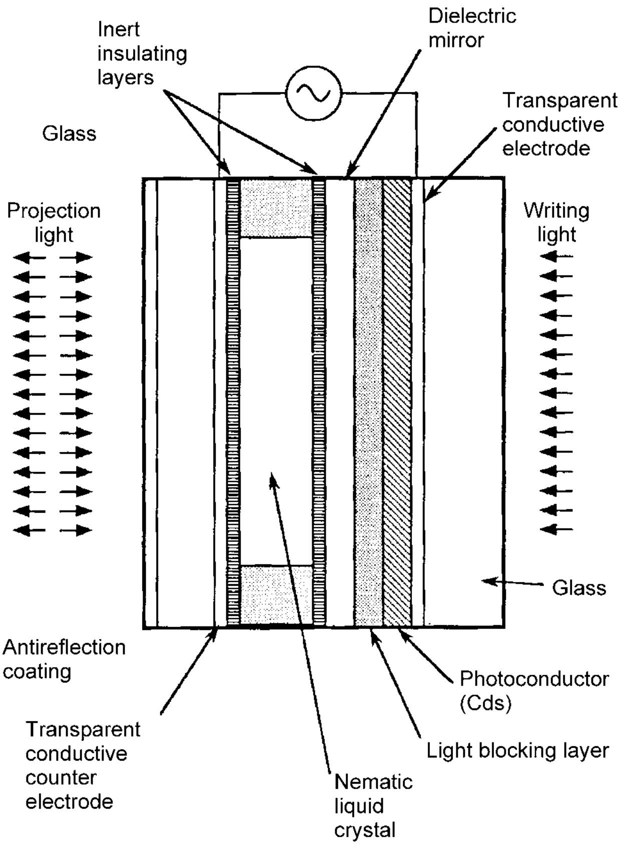

Inert
insulating
layers
Dielectric
mirror
Transparent
conductive
electrode
Glass
Projection
light
Writing
light
Glass
Antireflection
coating
Photoconductor
(Cds)
Transparent
conductive
counter
electrode
Light blocking layer
Nematic
liquid
crystal

Figure C2.4.3. The Hughes PA LCLV.

of the LC layer results in a change in the polarization rotation of the LC leading, with the use of a polarizing beam splitter, to an intensity modulation of the out-going optical beam.

In the particular case of the Hughes LCLV device shown in figure C2.4.3, the two LC configurations commonly used are the hybrid-field-effect mode using a 45° TN configuration [57] or, alternatively, the tilted-perpendicular mode based on the controlled-birefringence effect in a vertically aligned, negative-anisotropy LC, which has been in use more recently [113]. These devices initially employed CdS as the photoconductive layer which has been replaced by an amorphous-silicon layer. A fast-response version of the device for colour-sequential mode operation based on single-crystalline silicon photosubstrate was also demonstrated [114]. Due to the simplicity of an electrically addressed system on the one hand, and the rapid improvement in the yield of large CMOS arrays on the other, the trend today is towards electrically addressed LCD systems for projection displays. However, high-end applications requiring exceedingly large array size ($>2000 \times 2000$) and high brightness (>5000 lm) are still a hard reach for electrically addressed LCDs. The 'old-time' PA LCLVs are therefore still potential candidates for such demanding projection display applications such as electronic cinema [113]. In addition to reflective-mode devices, transmissive-mode PA SLMs have also been demonstrated [115–117], using a variety of semi-transparent photoconductors including amorphous silicon and BSO. The latter was developed at the Thales Research Laboratories, under Professor J-P Huignard. These devices were constructed using nematic, FLC and PDLC LC-modulators.

The typical performance of a PA a-Si LCLV is given in table C2.4.6.

Table C2.4.6. Performance parameters of the PA, a-Si LCLV.

Performance parameter	Description/value	Units
Device configuration	Reflective-mode	N/A
Optical modulator	Nematic liquid crystal	N/A
Driver array	Amorphous Si-photo-substrate	N/A
Array size	2000×2000–equivalent	Pixels
Resolution	25	LP mm^{-1} @ 20% MTF
Temporal response	Rise/decay time: 0.5–10 ms, frame-rate: 30–300 Hz	ms, Hz
Spectral bandwidth	Input beam: 550–700, output beam: 450–650	200 nm
Input sensitivity (PA device)	0.1–0.3	$\mathrm{mW\,cm}^{-2}$
Contrast ratio	200:1 (40 dB)	dB
Output modulation efficiency	10–30 (unpolarized, collimated beam)	%
Output uniformity	1–3	%
Physical array size	Diameter ~50	mm
Power consumption	~100	mW
Operating temperature range	10–40	°C
Shock resistance	1–5	G

C2.4.3.2 MEMS-based SLMs

General

The origin of these devices occurred after various attempts in the late 1970s to use continuous membrane mirrors, locally deformed under the application of local electric fields, as a means to modulate the reflectivity or the phase of the reflected optical beam [118].This concept was expanded in the early 1980s to include the first MEMS-type structures in silicon [119, 120]. This MEMS structure is essentially an array of silicon-based pixel-level metallized cantilevers which tilt in response to an electric potential applied to the pixel electrode. This MEMS structure which started with analogue tilt response design was later refined by the Texas instruments developers as a binary (digital) device with only two states. This recent design, combined with a complex driving circuitry to allow grey scale operation based on a PWM scheme, is the basis for today's high-brightness, high-resolution projectors aimed at such ambitious goals as electronic cinema [121]. The other principal MEMS-type SLM is based on the pixels constructed in a shape of interdigitated fingers. In this configuration, the application of an electric field across this digitated finger structure results in the formation of a diffraction grating with a variable field-dependent depth. Thus diffraction of the incoming optical readout beam, rather than its deflection, constitutes the novelty of this MEMS structure. This 'grating light valve' (GLV) device [122, 123], conceived and developed by Bloom, is now under development.

Finally, another recent related development has been reported, which is developing a similar structure of a MEMS-silicon-based micro-mirror SLM. However, rather than using an electrostatic field to drive the micro-mirrors, the device uses a piezo-electric micro-transducer fabricated in each pixel [124]. Finally, for completeness, we should also mention a recent effort by researchers to develop mechanical continuous membrane implementations of SLM similar in concepts to the earlier attempts mentioned [125] above.

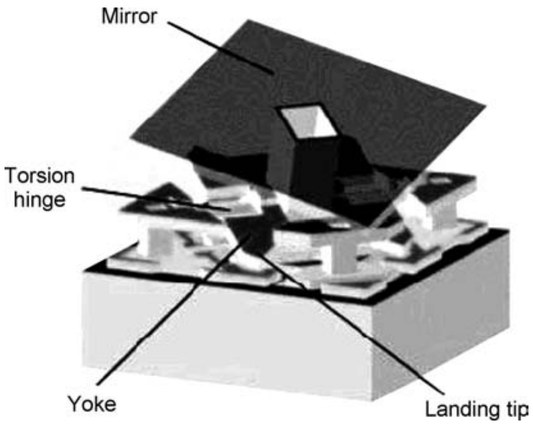

Mirror
Torsion hinge
Yoke
Landing tip

Figure C2.4.4. The pixel structure of the digital micro-mirror device. (After [121].)

The micro-mirror SLM (digital micro-mirror device) [120, 121]

The digital mcro-mirror device (DMD) SLM (figure C2.4.4) is based on a MEMS structure that is fabricated using a CMOS-compatible processes over a driver array which is based on a CMOS memory.

Each pixel modulator consists of a $16 \times 16\ \mu m^2$ aluminium mirror, which can reflect light in one of the two directions depending on the state of the underlying driver circuit. In the on-state, the mirror is rotated to: $\theta_L = 0°$ (into the system's FOV). In the off-state, the mirror swings to: $\theta_L = -10°$ (i.e. out of the system's FOV cone). Thus, by using the DMD in conjunction with a light source and a suitable projection optics (figure C2.4.5), the mirror reflects incident light either into or out of the pupil of the projection lens.

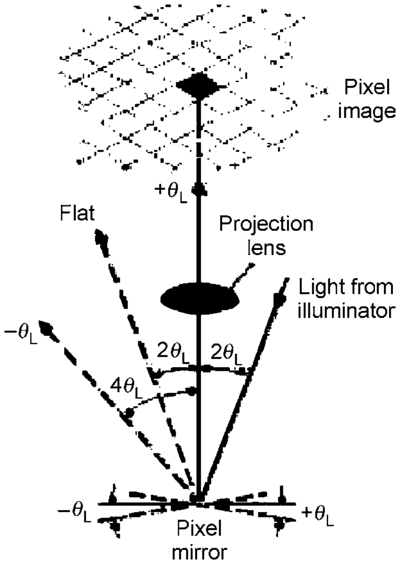

Pixel image
$+\theta_L$
Flat
Projection lens
Light from illuminator
$-\theta_L$
$2\theta_L$
$2\theta_L$
$4\theta_L$
$-\theta_L$
Pixel mirror
$+\theta_L$

Figure C2.4.5. The optical system arrangement for the DMD modulator. (After [121].)

Thus, with the above arrangement, the on-state ($\theta_L = 0°$) of the mirror appears bright, while the off-state of the mirror ($\theta_L = -10°$) appears dark.

The DMD pixel mirror is constructed over a CMOS driver array similar to an SRAM cell. An air-gap between the driver circuitry and the mirror is formed using an organic sacrificial layer. The air-gap allows the mirrors to rotate about two flexible torsional hinges. The mirror is rigidly attached to an underlying yoke, which, in turn, is connected by two flexible torsional hinges to support posts that are formed on the underlying substrate.

An electrostatic field generated by the underlying pixel results in a mechanical torque applied on the micro-mirror. This torque, applied against the restoring torque of the hinges, produces a mirror rotation in the positive or negative direction. The mirror and yoke rotate up to the point that the yoke comes to rest against the mechanical stops.

The DMD mirrors are 16 μm^2 made of aluminium for high reflectivity. They are arrayed to form a matrix having a high fill factor (approximately 90%) to attain high optical throughputs.

Grey scale is achieved by using a PWM technique. This is an addressing method in which the effective frame rate is M times (e.g. $M = 256$) faster than the actual, visual frame rate (e.g. 30 Hz). The binary-modulating pixels are activated in duty cycle ratios between $1/M$ amd 1.0 during the modulation of each visible frame. This, combined with the integrating response of the eye, creates an effective grey scale perception. Colour operation is achieved by using colour filters.

The system can also operate in a colour-sequential mode using sequentially flipped colour filters, in conjunction with a single device. Both the PWM grey scale method and the colour-sequential operation are possible due to the relatively fast electro-mechanical response of the micro-mirror which features typical switching times of ~20 μs. Thus, assuming a 3 × 8-bit operation for a 256-level RGB, colour-sequential video projection, the pixel switching requirements are: $T_{\text{switch}} \sim (16\,\text{ms})/[(256\,\text{levels/colour}) \times (3\,\text{colours})] \sim 21\,\mu\text{s}$, which can still be met by the ~20 μs switching time of the micro-mirror.

As to the reliability issues related to the mechanical failure of the micro-mirror, due to the continuous flipping, TI reports that testing of hinge fatigue resulted in over 1×10^{12} (1 trillion) cycles without mechanical failure. This 20 year equivalent operation, performed in an accelerated cycling test, seems to indicate that hinge fatigue is not a reliability concern for the life of an ordinary DMD product.

In the brighter, colour-parallel mode, two or three DMD arrays are used with stationary colour filters to produce a full colour image table (see table C2.4.7).

The grating light valve [122, 123]

The GLV technology, originally developed by Bloom, similar to the DMD, is based on the MEMS techniques, to form pixels in a silicon chip. Each of these pixels (typically, 25 μm size) is made up of multiple ribbon-like structures (figure C2.4.6), which can actually be moved up or down over a very small distance (only a fraction of the wavelength of light) by controlling electrostatic forces. The main advantage of the GLV-MEMS technology over the DMD-MEMS technology is in the reduced displacement required of the strips (1/4 λ or around 0.15 μm), relative to the DMD pixel displacement (around ±1.5 μm for the pixel tip). This leads to a substantially faster response speed of the order of 1 MHz or better.

The ribbons are arranged in such a way that each pixel is capable of either reflecting or diffracting light. An output image is formed by collecting the reflected or diffracted light with an appropriate lens system, either projected onto a front- or rear-screen system, or to be viewed directly by the eye. A Schlieren optical system is used to discriminate between the two optical states. By blocking reflected light and collecting diffracted light, contrast ratios of a few hundreds

Table C2.4.7. Summary of the DMD-SLM performance.

Performance parameter	Description	Units
Input/output	Electrically addressed, readout beam, intensity modulated	N/A
Device configuration	Reflective mode	N/A
Optical modulator	Micro-electro-mechanical mirror	N/A
Driver array	CMOS driver array	N/A
Array size	Up to 1920 × 1080	Pixels
Resolution	25 LP mm^{-1} @ 50% MTF	LP mm^{-1} @ MTF
Temporal response	Opt. switching time: 20 μs; frame-rate: 30–300 Hz (colour, 8-bit); 10 kHz binary (with custom addressing)	ms, Hz
Spectral bandwidth	Output beam: 450–650	200 nm
Input sensitivity	N/A	N/A
Contrast ratio	100:1 (20 dB)	dB
Output diffraction efficiency	60	%
Output uniformity	1	%
Physical array size	~40 × 20	mm
Power consumption	~1	W

to one can be achieved. In an ideal square-well diffraction grating, 81% of the diffracted light energy is directed into the ±1st orders. By adding multiple Schlieren stops and collecting more orders, practical systems can achieve greater than 90% diffraction efficiency (DE). The gaps between GLV ribbons (defined by the minimum lithographic feature) control the optical diffraction efficiency. Thus, the theoretical DE varies between 82 and 98% for ribbon gaps between 1.2 and 0.35 μm, respectively.

The GLV technology can be employed in either digital or analogue modes. In the digital addressing mode, the switching is based on the PWM described earlier for the DMD device operation. In the analogue mode, the depth to which ribbons are deflected is controlled by the driving circuitry. When

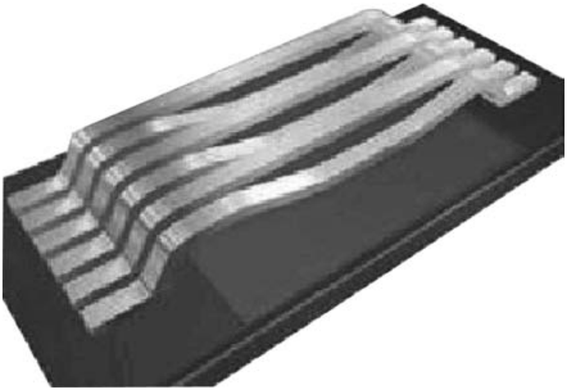


Figure C2.4.6. The pixel structure of a GLV device.

the ribbons are not activated (deflected), the pixel is in its off-state. When the ribbons are deflected to (1/4) λ, the pixel is fully on. Any particular grey level can be generated by deflecting the ribbons to positions between these two limits.

For display applications, the GLV is operated in the scanned mode, where a linear array of GLV pixels is used to project a single vertical column of the image data. This column is optically scanned at a high rate to produce a complete 2D image. As the scan moves horizontally, GLV pixels change states to represent successive columns of video data, forming one complete 2D image per scan. The relatively fast switching speed of GLV devices (of the order of 1 MHz) allows full HDTV scanning (1920 × 1080 image) at video rates of up to 96 Hz.

The main performance parameters of the device are summarized in Table C2.4.8.

The thin-film micro-mirror array [124]

A thin-film micro-mirror array (TMA) display is currently under development by researchers at Daewoo Electronics in South Korea.

The modulator technology of the TMA device is similar to that of the TI DMD discussed earlier. However, rather than using electrically activated micro-cantilevers, it uses micromachined thin-film piezoelectric actuators to control the micro-mirror tilt mechanism in an analogue fashion, thus enabling grey scale operation with over 256 levels. Each pixel consists of a mirror and an actuator. In previous designs, the two had been co-planar, but the improved design has the actuator situated below the mirror, increasing the fill factor up to 94% and the contrast ratio to 200:1.

The TMA uses thin-film piezoelectric actuators in the form of micro-cantilevers, which consist of a supporting layer, bottom electrode, piezoelectric layer and a top electrode.

Table C2.4.8. Performance parameters of the grating light valve.

Performance parameter	Description/value	Units
Input/output	Electrically addressed, readout beam, intensity modulated	N/A
Device configuration	Reflective/diffraction mode (1D)	N/A
Optical modulator	Micro-electro-mechanical grating	N/A
Driver array	CMOS driver array	N/A
Array size	1D: 1920 × 1 (2D also possible)	Pixels
Resolution	25	LP mm^{-1} @ 50% MTF
Temporal response	Opt. switching time: down to 20 ns; line-rate: 30–300 Hz (colour, 8-bit); 10 kHz binary (with custom addressing)	ns, Hz
Spectral bandwidth	Output beam: 450–650.	200 nm
Input sensitivity	N/A	N/A
Contrast ratio	>200:1 (40 dB)	dB
Output diffraction efficiency	Up to 80	%
Output uniformity	1	%
Physical array size	~50 × 0.1 (for 25 × 100 μm pixels)	mm
Power consumption	<100	mW

The micro-mechanical slit-positioning SLM [126]

This is a recent concept based on the technology of MEMS described earlier, where an actuator can programmably shift an optical slit of width between 8 and 100 μm, thus enabling the position of the transmitted light regions to be changed as well as modulate the incoming light beam.

The proposed SLM architecture can be 1D or 2D (for imaging). The applications include micro-spectrometer systems as well as the next generation space telescope with multi-object spectrometer under development by the European and US space agencies.

When an electric field is applied to the two electrodes, the mechanical strain in the piezoelectric layer causes a vertical deflection of the mirror. The response time of each pixel is 25 μs, making it fast enough for field-sequential colour display applications. The display, which is used to make a high-brightness XGA-format projector, has an optical efficiency of 20% at a panel size of 2.54 inches.

In order to modulate the light intensity of the individual mirror pixels projected on the screen, a field-stop is used as a light valve. When a mirror does not tilt, all the light reflected by the mirror is blocked by the field stop and the pixel is at its off-state (dark).

When the mirror is fully tilted, all the light goes out through the projection stop and the pixel is at its brightest state (white).

C2.4.3.3 MQW modulators

Introduction

While this is the most recently developed SLM technology, it is also the most promising one for demanding, ultra-fast-frame-rate (>MHz) applications such as optical data processing. It was pioneered in the early 1980s by Miller, who demonstrated, for the first time, the potential of artificially made stacks of quantum-size layers of alternating GaAs (quantum well material) and GaAlAs (quantum barrier materials) for a highly efficient, ultra-fast electro-absorption (EA) effect [127, 128] with potential use for very fast SLMs. The effect was named the 'quantum confined stark effect' (QCSE) by Miller. The interest at that time was that of fast switch or switch arrays for communications. While the EA effect can be observed in direct-gap semiconductors such as GaAs and InP (the so-called Franz–Keldysh effect), it is relatively inefficient in requiring high-voltage switching to accomplish a relatively poor contrast modulation. In fact, an attempt to use this effect in constructing a CCD-driver-based, GaAs-SLM was attempted in the early 1980s [129]. The low contrast accomplished (~1.2:1) convinced the technical community that this is not the right path to a fast, efficient SLM technology. The subsequent development of MQW-based SLM technology [130] and the, additional use of Fabry–Perot structure subsequent, combining the EA effect with the related electro-refraction (ER) effect, resulted in an SLM technology capable of gigahertz response with contrast ratio of 100:1 and higher, at moderate addressing voltage levels of ~5 V. This SLM technology constitutes the only potential solution known today for the demanding optical data processing applications.

MQW modulators: physical background

The MQW structure, as its name implies, consists of a stack of quantum wells (QWs) i.e. a few molecular monolayers of a low-energy-gap 'well' structure (e.g. GaAs), sandwiched between thin wide-bandgap 'barrier' layers (e.g. GaAlAs). This structure 'compresses' the excitonic wave function (bound electron–hole pair) due to the ultra-thin dimension of the well layer in which the excitons reside, forcing the electron–hole pair to be much closer to each other. This in turn, results in a much higher coulombic energy and, therefore, higher ionization energy for the excitons compared to their

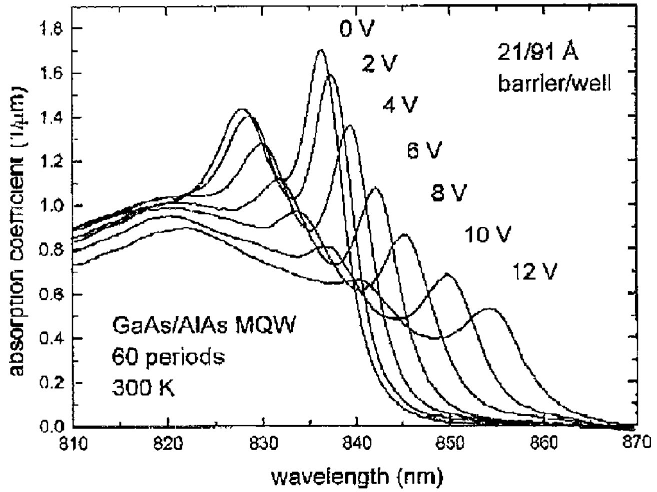


Figure C2.4.7. Field-induced absorption changes in an MQW sample. (After Goossen *et al* 1994 *Appl. Phys. Lett.* **64** 1071.)

ionization energies in a bulk semiconductor material. This allows excitons to exist in these QWs at room temperatures. Furthermore, the application of an electric field of the order of $100\,\mathrm{kV\,cm^{-1}}$ causes a partial separation of the electron–hole wave functions, thereby altering the energy levels of the excitons. This, in turn, alters the effective optical resonance of the exciton thereby shifting the resonant absorption wavelength. Since the wavelength dependence of the absorption curve is quite steep near the resonant wavelength, this shift can effectively be used to modulate the optical absorption of the optical beam (figure C2.4.7). The importance of this effect for SLM application is the fact that such modulation can be effectively performed with only a few volts of bias resulting in an EA effect with picosecond response times.

The theoretical treatment of the EA effect in a QW predicts a quadratic effect of the resonant wavelength dependence upon the applied electric field [131]. The resonant excitonic EA effect in the MQW structure is expected to result in a corresponding resonant ER effect through the Kramers–Kronig relationship which links the spectral dependence of both the absorption coefficient and the refractive index [132].

Another important type of MQW structure, the compositionally doped 'nipi' structure, was initiated and developed by Ploog and Doehler [133, 134]. This structure is based on alternating compositionally doped quantum-size layers, e.g. n-GaAs/p-GaAs. Often an insulating layer ('i'-layer) is inserted between the n- and p-layers and hence the name 'n–i–p–i'. The resulting stack of quantum-size p–n junctions, when selectively contacted with all the n- and the p-layers in parallel, can result in a very large EA or ER effect [135].

Finally, in reviewing related structures and physical effects of MQW, lower-dimensionality quantum well structures should also be mentioned. The formation of quantum well structures results in the confinement of carriers to 2D 'wells', as pointed out earlier. From the standpoint of the dynamic behaviour, such confinement results in the quantization of the density of energy states which now behave as a series of step functions. This effect, which actually contributes to the steep spectral absorption curve and hence to the enhancement of the EA, can be enhanced even more by the confinement of the carriers to a single dimension ('quantum lines') or ultimately to zero-dimensional 'quantum dots'. In particular, quantum dot structure has been extensively researched in the last few years due to the theoretical

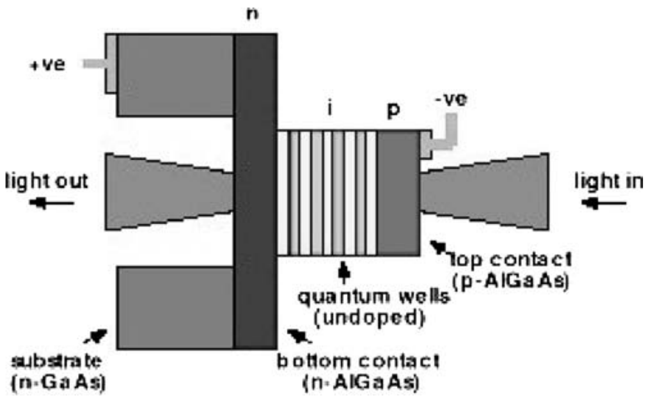


Figure C2.4.8. MQW modulator (After Miller D A B 1990 *Int. J. High Speed Electron.* **1** 19–46.)

potential of yielding extremely sharp delta-shaped absorption curves. These could result, in turn, in an extremely effective switching and therefore in very efficient SLM structures. The main issue in the manufacturing of these structures has been the finite spread in the dot size distribution, which sharply reduces the steepness of the absorption curve [136].

MQW modulator: device structure

The basic device structure of an electro-absorbing MQW modulator in a transmissive mode is shown in figure C2.4.8. The MQW structure is fabricated (using either MBE or MOCVD epitaxial growth techniques) within a p–i–n structure. This configuration allows the generation of a relatively high electric field created by back-biasing the p–n junction and extending into the low-doped, high-resistivity intrinsic (I) region where the MQW structure is situated.

The medium–high-doped n- and p-AlGaAs regions constitute, in effect, transparent conductive electrodes, as these AlGaAs-based layers have a wider bandgap than the QW excitonic line and, at the same time, feature relatively high sheet-conductivity due to their high doping levels.

The transmission contrast ratio of an MQW sample can be roughly estimated as CR $\sim \exp(\Delta\alpha L)$, where $\Delta\alpha$ is the field-induced (negative) change in the excitonic absorption coefficient, and L is the optical path length or the MQW sample's thickness.

As can be seen from figure C2.4.7, an absorption coefficient change of: $\Delta\alpha \sim (-)1\ \mu\text{m}^{-1}$ is induced for a voltage change of approximately 10 V over a sample thickness of approximately 0.7 μm, which translates to a field variation of approximately $\sim 1.4 \times 10^5\ \text{V cm}^{-1}$. Based on the estimated CR above, this absorption change translates to a contrast ratio of around 2:1. This is a typical value for the performance of a transmissive-type MQW device, having around 50 QW periods, with around 10 V of applied voltage bias. For a reflective-type device, one can theoretically attain up to a quadratic enhancement of the CR since the optical path L in the CR equation above is doubled.

This translates to around 5:1 in reflectivity contrast for modulators operating in a reflection mode. To go beyond these figures necessitates either a large QW stack, (e.g. 200 periods allows over 10:1 contrast in transmission [130]) or other means of amplifying the quantum-confined Stark effect. In the following, we will shortly describe two such configurations. Historically, the first attempt was to use the photo-current generated inside the MQW modulator to affect the device's bias voltage in a positive feed-back mechanism, so as to enhance nonlinearly the QCSE by having the voltage bias of the sample

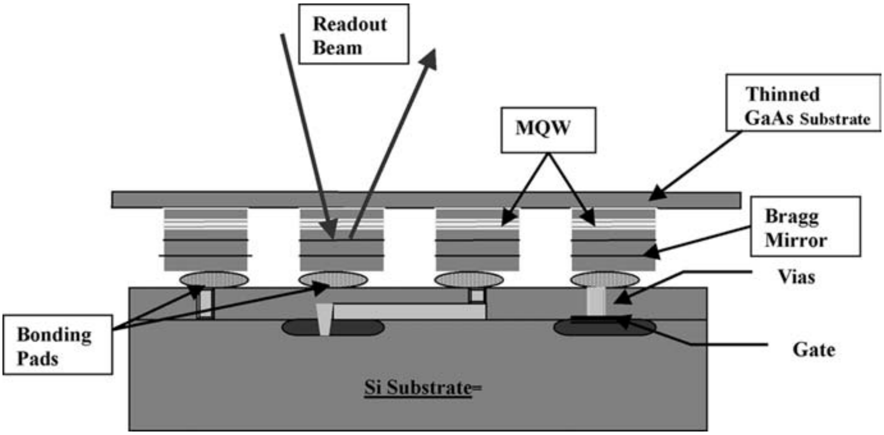


Figure C2.4.9. Schematics of a Fabry–Perot MQW-SLM [154].

be varied in proportion to the photocurrent. These early self-electro-optic effect devices (SEEDs) [137] were designed to get a low-field nonlinear, bi-stable behaviour in PIN-type MQW devices.

A switching energy as low as 180 pJ for a 60 μm × 60 μm SEED has been demonstrated [138]. A second, common modification of the QW device is the integration of the device within an optical resonator cavity. This essentially 'amplifies' the EA and/or ER effect by effectively extending the optical path via the multi-pass effect of the cavity.

The asymmetric Fabry–Perot QW modulator [138–140] is based on the insertion of the MQW stack within an optical resonant cavity composed of a semi-transparent top mirror (figure C2.4.9) and a bottom, high-reflectivity, 1/4λ-stack dielectric mirror. In this case, the resonator's reflectivity, R_R, is given by: $R_R = [R_T(1 - R_\alpha/R_T)^2/(1 - R_\alpha)^2]$, where $R_\alpha = \sqrt{R_T R_B}\exp(-\alpha d)$ and R_T, R_B, are the reflectivities of the top and bottom mirrors, respectively.

As can be seen, if the effective bottom-mirror reflectivity ($R_B \exp[-2\alpha(E)d]$) is sufficiently reduced by the field-induced absorption $\alpha(E)$, to the point where it matches the top-mirror reflectivity: $R_B \exp(-2\alpha(E)d) = R_T$, we have a *zero reflectivity* for the resonator. This capability of bringing down the reflectivity close to a zero level implies that we can expect high reflectivity contrast ratios for the relatively modest bias levels. The above estimate did not take into account the ER effect that must also be considered [138].

Contrast ratios in excess of 100:1 were obtained for moderate sized QW stacks and biases lower than 10 V [140]. It should be pointed out, however, that such use of a thickness-sensitive, resonant cavity makes it difficult to accomplish adequate spatial uniformity of the modulator array.

Phase modulation using MQW

An earlier attempt to demonstrate phase modulation in MQW was reported in 1988 [141]. The demonstration of 0.21 @ 852.5 nm was accompanied as expected by significant attenuation. As pointed out earlier, the attenuation is expected as a direct result of the Kramers–Kronig relationship. Thus, since the optical modulation effect in an MQW modulator is based on the exitonic resonance effect, one expects the maximum ER effect to always occur in the vicinity of the absorption peak. To help to maximize the ER effect while minimizing the absorption, one can define a figure of merit by: $\mathrm{FM} = \Delta n(E)/\alpha(E)\lambda$, which can be used to search for material systems (and/or spectral regions) with a high ratio of refractive to absorptive optical modulation [142, 143]. This formulation implies that operation at a shorter-wavelength region may hold advantage for this type of modulation. However, earlier attempts to use this expression in search of material systems exhibiting significantly high FM were not too encouraging [142]. A more recent attempt indicates a possibility of performing phase modulation for beam-steering applications [144]. However, the authors do not indicate the absorption-associated insertion losses in this device.

MQW SLMs

Earlier attempts to construct small-size arrays of MQW-SLM devices were made [145] (6 × 6 array) and [146] (128 × 1 array).

More recent attempts of constructing and demonstrating larger-size arrays of MQW-SLMs were made [147] (128 × 128 array) and the successive effort [148, 149] (256 × 256 array). The latter was reportedly operated at up to 300 000 frames per second, allowing up to 600 000 correlations per second to be performed on a 128 × 128 array. The device was based on GaAs/GaAlAs MQW stacks in a PIN geometry, inserted within an asymmetric FP cavity discussed earlier, with pixel sizes of around 40 μm × 40 μm. The group claimed up to 6 bits of grey levels achievable on their 2048 × 1 1D devices. Summary of the MQW-SLM performance is given in table C2.4.9 below.

The demonstration of a 16 × 16 cross-bar switch ‘Amoeba’, for optical communications was recently demonstrated [150]. The switch is based on flip-chip bonding of a 0.8 μm technology CMOS chip to an MQW array of detectors/modulators.

Optically addressed MQW-SLMs were also conceived [138] and demonstrated [151].

The recent demonstration [151] is based on a reflective-mode, GaAs/GaAlAs MQW modulator constructed inside a Fabry–Perot cavity, and combined with a free-carrier trapping layer. The latter is sensitive to the incident, near-IR to visible optical beam. A near-excitonic resonance probe beam reads out the spatial optical signal-modulated reflectivity of the Fabry–Perot cavity to form an

Table C2.4.9. The main performance parameters of a state-of-the-art, electrically addressed MQW-SLM.

Performance parameter	Description	Units
Input/output	Electrically addressed, readout beam, intensity modulated	N/A
Device configuration	Reflective-mode	N/A
Optical modulator	MQW-electro-absorption mode	N/A
Driver array	CMOS driver array	N/A
Array size	256 × 256	Pixels
Resolution	(est.) 25 LP mm^{-1} @ 50% MTF	LP mm^{-1} @ MTF
Temporal response	Opt. switching time: ~10 ps; frame-rate (CMOS limited): 300 kHz	ps, kHz
Spectral bandwidth	Output beam centre-λ ~850, bandwidth ~5	nm
Input sensitivity	N/A	N/A
Contrast ratio/grey levels	100:1 (20 dB)/ 6 bits	dB
Output modulation efficiency	~30	%
Output Uniformity	1	%
Physical array size	~10 × 10	mm

Based on [148, 149].

output image. Other efforts in the development of MQW-SLM arrays that should be mentioned are [152] in developing GaInAs-based MQW modulator arrays for cross-bar switch applications.

A recent effort is based on MQW modulator-array prototypes fabricated using the GaAs/AlGaAs material system. The first prototypes fabricated are designed with an array size of 128 × 128 pixels, made with a pixel pitch of 38 μm [153].

Finally, significant improvements in the uniformity of FP-based MQW arrays (figure C2.4.9), reducing optical nonuniformity to 3.3 nm across the 4 inch wafer, have recently been demonstrated [154]. These devices operate at low voltage (2.1–3.9 V), producing a reasonably high contrast ratio (16:1–98:1) and are designed for a frame rate of up 300 kHz. The array was hybridized to an 8-bit 0.25 CMOS technology, Si driver.

C2.4.3.4 Solid-state electro-optic SLMs

In general, there has not been much progress in solid-state electro-optic SLMs. These devices, which constituted the 'first-generation' SLMs in the late 1960s to early 1980s, were quite complex and bulky due to the very high bias voltage required, and the related large pixel size/SLM apertures required (see discussion on the fundamental limitations of SLMs in section C2.4.6 later). Thus, devices based on KD*P and $LiNbO_3$, turned out to be commercially unviable for the main display applications, and gave way to LCDs as the SLM technology of choice for displays and to some extent for optical processing. Some references to these technologies can be found in the earlier conferences on SLM technologies [155–157].

One electro-optic SLM that should be mentioned is the PLZT-based device. There have been recent efforts in trying to form 1D and 2D products using this technology [158]. The interest in this modulator technology is in the relatively high electro-optic coefficient with half-wave voltages of the order of 100 V [159, 160].

C2.4.3.5 Acousto-optic SLMs

The acousto-optic (AO) modulator technology has been largely implemented in the form of either bulk devices such as the Bragg cell or as surface acousto-optic (SAW) devices. This is a fairly mature technology of 1D modulator arrays which has been in existence since the early 1930s [161].

The general structure of an AO modulator is shown in figure C2.4.10. A piezo-electric transducer modulated by RF (100–3000 MHz) generates an acoustic wave that propagates within the AO crystal (e.g. quartz). The propagating density grating formed in the crystal gives rise to a corresponding, propagating refractive index grating. The latter diffracts an optical beam launched at the AO crystal perpendicular to the grating. The grating constant, which is essentially the acoustic wavelength Λ, varies inversely with the modulating RF frequency f, i.e. $\Lambda = C_s/f$, where C_s is the acoustic (sound-wave) velocity in the material. The Bragg diffraction angle θ, which is inversely proportional to the acoustical grating period, can be varied by modulating the RF frequency of the piezoelectric transducer. This device can therefore deflect an incident optical beam into a range of diffraction spots thereby acting as a 1D SLM.

A typical performance parameter of an AO Bragg cell is given in table C.2.4.10. From the standpoint of the SLM technology, a key parameter specifying the number of available spots or 'pixels' is the time–bandwidth product. This parameter is roughly the product of the frequency bandwidth (200 MHz in this case), and the time aperture (5 μs), which is determined by the 'time of flight' of an acoustic wave ($C_s = 3.63\ \mathrm{mm}\ \mu\mathrm{s}^{-1}$) across the crystal length (19 mm), giving around 1000 resolved spots.

For further information on the AO technology, we refer the reader to several extensive technology surveys which have been recently published [162–164].

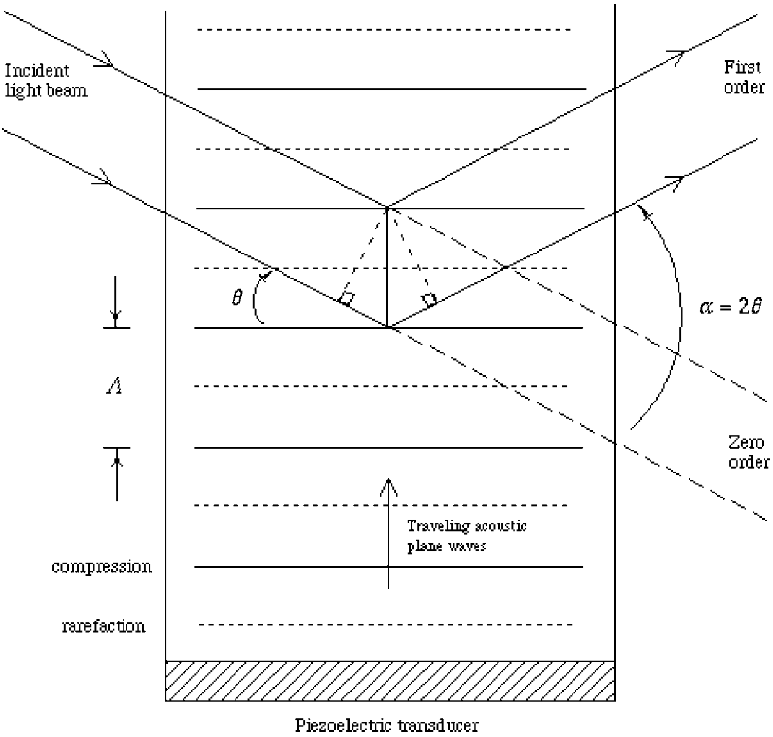

Incident light beam
First order
θ
α = 2θ
Λ
Zero order
Traveling acoustic plane waves
compression
rarefaction
Piezoelectric transducer

Figure C2.4.10. Schematic of the acousto-optic effect.

C2.4.3.6 Magneto-optic SLMs

The employment of the magneto-optic effect for spatial modulation is shown in figure C2.4.11 [165–167]. The device, which is based on the magneto-optic effect associated with the magnetic domain

Table C2.4.10. Typical performance parameters of an AO Bragg cell[a].

Operating wavelength	Any within the range 442–850 nm
Time–bandwidth product	1000
Centre frequency	300 MHz
3 dB bandwidth	200 MHz
Active aperture	1.5 mm H × 19 mm L
Time aperture	5 μs
Interaction medium	$PbMoO_4$
Acoustic velocity	3.63 mm μs^{-1}
Diffraction efficiency	10% at 1 W RF power (633 nm)
Electrode	Apodized to minimize acoustic walkoff
Optical surface flatness	wavelength/10 or better
Input impedance	50 Ω
Input VSWR	<2:1 across RF bandwidth
Optical reflectivity	<5%/surface

[a] Courtesy of Isomet Co.

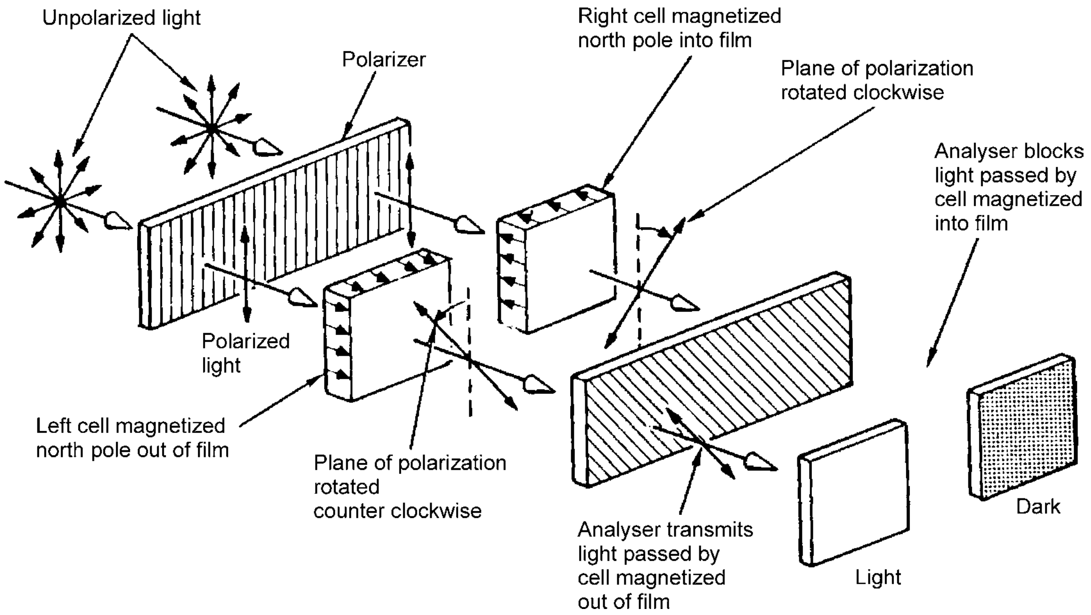

Unpolarized light
Polarizer
Right cell magnetized north pole into film
Plane of polarization rotated clockwise
Analyser blocks light passed by cell magnetized into film
Polarized light
Left cell magnetized north pole out of film
Plane of polarization rotated counter clockwise
Analyser transmits light passed by cell magnetized out of film
Light
Dark

Figure C2.4.11. The magneto-optic effect for spatial light modulation. (After [165].)

reversal in a ferromagnetic material (gadolinium–gallium–garnet (GGG)), is electrically driven using conducting X–Y mesh-lines which form the boundaries of the magneto-optical pixels. As a particular mesh-node is cross-activated (i.e. both the X-line and the Y-line for that node) the resulting current flowing around that pixel corner generates a magnetic field, which causes a magnetic domain reversal in that magneto-optical pixel.

Although initially developed as a binary (phase or amplitude) SLM, the presence of a 'neutral' state of the magnetic domains later allowed the realization of a ternary-state device [166]. Due to its relatively high insertion losses in most of the visible spectrum, the device was targeted mainly for optical processing applications. The relatively high current required to switch to the on-state is one of the main drawbacks of this device, effectively limiting its frame rate to the low kilohertz range (see table C2.4.11). Recently, a renewed effort in the development of the MO-SLM was reported [168, 169]. An overall improvement by a factor of three in power consumption using a modified shape driveline with detailed device simulation was reported.

C2.4.3.7 Photorefractive SLMs

This modulation technology is based on the refractive index modulation in response to an incident optical beam. The effect has been extensively studied in the past 30 years [170–173] and, hence, only a brief description will be presented. Figure C2.4.12 describes the process of photo-refraction upon the illumination by an interference pattern generated by the illumination of two coherent beams (figure C2.4.12(*a*)). The spatial periodic illumination results in the generation of periodic, local photo-carriers (e.g. electrons—figure C2.4.12(*b*)). These drift apart due to diffusion resulting in the formation of a space-charge field (figures C2.4.12(*c*) and (*d*)). Finally, the generated space-charge field results in a periodic modulation of the refractive index due to the electro-optic effect in this material. The resulting diffraction efficiency is therefore proportional to the appropriate linear electro-optic coefficient r, with a figure of merit for the material which can be written as [174]: $Q = n^3 r/\varepsilon$, where ε is the dielectric

Table C2.4.11. Summary of the magneto-optic SLM performance parameters.

Performance parameter	Description	Units
Input/output	Electrically addressed, readout beam, intensity/phase modulated (ternary state: 0, ±1)	N/A
Device configuration	Transmissive/reflective modes	N/A
Optical modulator	Magneto-optic GGG	N/A
Driver array	CMOS driver array	N/A
Array size	128 × 128	Pixels
Resolution	Approximately 20 LP mm^{-1} @ 50% MTF	LP mm^{-1} @ MTF
Temporal response	Frame rate: 0.5–1.0	kHZ
Spectral bandwidth	630 (15%)–788 (45%)	nm
Input sensitivity	N/A	N/A
Contrast ratio/grey levels	>10 000:1 @ 788 nm	
Output modulation efficiency	15–45	%
Physical array size	~30 × 30	mm
Power consumption	~2–4	W

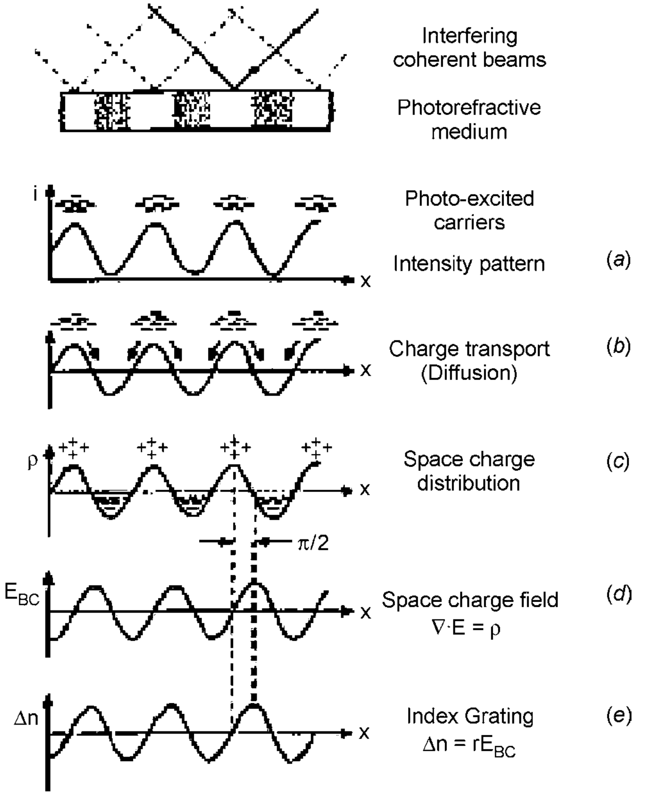

Interfering coherent beams
Photorefractive medium
i
Photo-excited carriers
Intensity pattern
x
(a)
x
Charge transport (Diffusion)
(b)
ρ
x
Space charge distribution
(c)
π/2
E$_{BC}$
x
Space charge field
∇·E = ρ
(d)
Δn
x
Index Grating
Δn = rE$_{BC}$
(e)

Figure C2.4.12. Schematics of formation of the photorefractive effect.

constant and n is the refractive index of the material. As it turns out, the response time of the effect is inversely proportional to the same figure of merit, Q [174].

While the effect has largely been studied for holographic storage applications, as well as adaptive optical corrections by phase conjugation (see section C2.4.5), the photo-refractive (PR) crystal can actually serve by itself, as a photo-addressed or PA SLM. Thus, with the formation of a photo-induced grating by the optical input signal, the device acts to modulate a probe (readout) beam. Since both the input and the output beams can contain spatially encoded information, the device actually acts as a PA SLM, although in this arrangement another input SLM is often used. This property of PR crystals has been used for numerous applications including phase conjugation [175], optical storage [176], photo-refractive correlators [177], laser beam cleaning [178, 179], information processing [179, 180] and novelty filter processors [181].

Early attempts to demonstrate spatial modulation included the photorefractive incoherent-to-coherent optical conversion (PICOC) device [182]. The process is based on modulating the pre-formed index grating generated in the photorefractive crystal by the interference of the two coherent beams, by the illumination of the incoherent (signal) image. Another demonstration of an incoherent-to-coherent spatial modulator was based on a grating-encoded phase modulation in a Ce-doped SBN crystal [183]. A more recent demonstration of this effect was based on the self-phase conjugation in an SBN crystal, in which a 28 lp mm^{-1} resolution of spatial modulation was demonstrated with a 6 mm × 12 mm crystal [184].

A more recent employment of the photorefractive effect was in the use of an optically addressed MQW-SLM by a US Naval Research Laboratory group [185]. The researchers developed an GaAs–AlGaAs, reflective-mode MQW device, shown in figure C2.4.13. This device is capable of operating both as a 'regular' PA SLM and as a photorefractive or holographic mode device. In this mode, two coherent pump beams generate a periodic photocharge pattern in the carrier absorption layer (figure C2.4.13(*b*)). The resulting space-charge field induces refractive index modulation within the MQW layers which in turn modulates the probe beam (figure C2.4.13(*f*)). Diffraction efficiencies of 1.5% were obtained with down to a 7 μm spot-size resolution @ $\lambda = 856$ nm. A principal advantage of this SLM technology is its very high spatial resolution, while the main drawback is its inherently slow response due to the energy

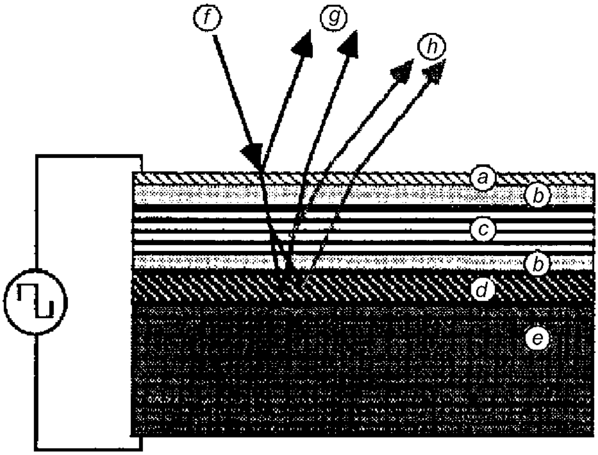


Figure C2.4.13. Photo-refractive-based MQW-SLM. (*a*) p-doped conductive layer; (*b*) free carrier trapping layer; (*c*) intrinsic MQW layer; (*d*) n-doped Bragg mirror; (*e*) n-doped GaAs wafer; (*f*) incident probe beam; (*g*) modulated reflected beams and (*h*) diffracted beams. (After [185].)

(time × power) to form the photo-refractive grating pattern, including the use of fast MQW configurations.

C2.4.3.8 Smart-pixel SLMs

A smart-pixel SLM (SPS) can be described as an SLM array consisting of opto-electronic pixel circuits where each of this circuits is capable of the following:

(1) Performing optical detection of the input beam.

(2) Performing some level of signal processing on the incoming signal.

(3) Optically modulating the output beam.

Such ‘smart pixel’ arrays clearly integrate some of the signal processing functions into the SLM itself [186, 187]. The main variable parameters of this SLM technology are:

(1) Detector array technology.

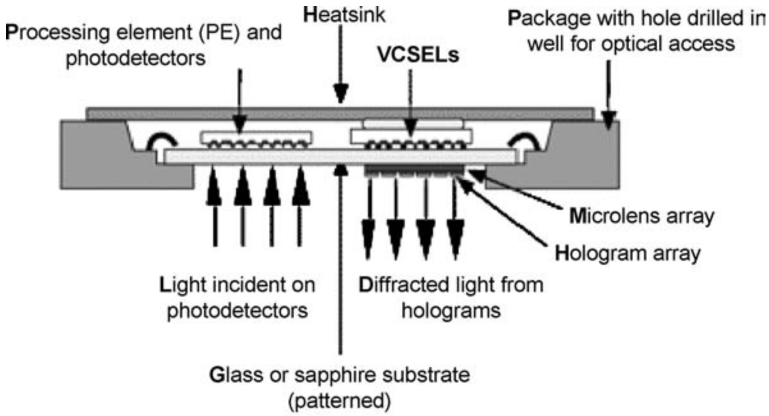
(2) Modulator array/modulated active emitter array technology.

(3) Signal processing functionality in each pixel.

Earlier concepts were based on silicon-PLZT modulators [188] where the silicon circuitry was used to provide both photo-detection and drivers for the PLZT electro-optic modulators. With the advent of liquid crystals and more recently QW modulators, smart SLMs were demonstrated with Si-FLC [189] as well as Si-VLSI MQW [187, 190] combinations. More recent configurations for SPSs involve the use of a combined Si-driver and VCSEL source/modulator [191], shown in figure C2.4.14.

In a recent development of a CMOS-based imaging smart-array system, a 2.5 MB s^{-1} bandwidth, over 30 dB in dynamic range with 150 μm smart pixels in an array size of 64 × 64 has been demonstrated [192].

Finally, the combination of both functions of imaging and display in a single SLM device has been demonstrated using CCD circuits operating both as an imager and a display [193]. A recent development


Processing element (PE) and photodetectors
Heatsink
VCSELs
Package with hole drilled in well for optical access
Microlens array
Hologram array
Light incident on photodetectors
Diffracted light from holograms
Glass or sapphire substrate (patterned)

The smart pixel array module. The original packaging for the VCSEL/Si scheme is dislayed above.

Figure C.2.4.14. VCSEL-Si-based smart pixel. (After [191].)

effort based on CMOS technology has been proposed with the combined imaging and display functions implemented in each pixel of the image transceiver array [194]. The application of this image transceiver device is targeted at the 'smart goggle'.

The main application for the SPS technology is the OI or cross-connects [187, 195–197]. Another important application is that of optical processing, where operations such as motion detection, pattern recognition, SIMD-parallel processing, neural net processing and analogue to digital conversion have been demonstrated [198–202].

C2.4.4 Novel devices and emerging technologies

The very successful application of SLMs in the booming area of displays, and the more recent interest in their potential use in the exploding communication field, has motivated the research and development of novel spatial optical modulator technologies and devices.

C2.4.4.1 Electro-holograms

This method is based on controlling the reconstruction process of volume holograms by means of an externally applied electric field [203–205]. The electro-holography (EH) effect exploits the voltage dependence of the photo-refractive effect in the para-electric phase of a material such as KLTN, which results in controlling the process of hologram reconstruction. Thus, the use of an electric field can result in the activation of pre-stored holograms in the photo-refractive material, which would otherwise appear optically homogeneous.

Such an effect when used in an array of such devices can be used to modulate, two-dimensionally, the optical intensity of a wave front or of an array of optical beams. Alternatively, it can be used to steer individually a 1D or 2D arrays of optical beams.

Furthermore, since this is essentially a controlled photo-refractive effect, the direction of the diffracted beam clearly depends on the beam's wavelength. This is of particular importance in the optical cross-connect (OXC) switch application for communications (e.g. for a WDM system) where one needs to have a wavelength-dependent, routing control of the optical beams. This topic is addressed in section C2.4.5 later.

Switching speeds of ~10 ns, and diffraction efficiencies of over 30% in 32 angularly multiplexed, volume holograms in a 3 mm × 3 mm × 3 mm KLTN crystal, which were switched on and off electrically, have been demonstrated.

C2.4.4.2 PBG devices

This is an extremely interesting optical-physics phenomenon introduced in the late 1980s. The main feature of photonic band-gap (PBG) structures [206, 207] is the periodic modulation of the refractive index n (or the dielectric constant), along one, two or three directions of space. In a composite formed by two dielectrics, the periodic modulation of one of the dielectrics effectively creates scattering centres, which are regularly arranged in the second medium, resulting in the coherent scattering of light. In this case, interference will eventually inhibit some frequencies that will not be allowed to propagate, thus giving rise to 'forbidden bands'. Under certain conditions, regions of frequency may appear to be forbidden regardless of the propagation direction in the PBG. In such a case, this region is said to present a full PBG. On the other hand, if the forbidden photonic band varies with the propagation direction in the region, a photonic pseudo-gap is spoken of. Furthermore, we can also introduce defects into the structure, resulting in the introduction of allowed energy levels into the gap, analogously to a doped semiconductor. All these facts permit us to establish a parallelism between the formalism used for electrons in ordinary crystals and that for photons in a photonic crystal (PX).

Continuing with this similarity, it is also possible to control the ‘optical conductivity’, i.e. the optical transmission of a photonic crystal, by modulating the relative indices of diffraction of the two materials of which the crystal is composed. In particular, if an electro-optical material, such as liquid crystal, is infiltrated into the optically periodic crystal (e.g. periodically etched, porous silicon), then controlling the effective refractive index of the LC via voltage bias can result in a significant shift of the forbidden band frequency. This in turn can drastically change the optical transmission. Hence, here we have another potential efficient candidate for SLM—since minute changes in the effective refractive index could lead to very significant variations in the optical transmission. This was actually demonstrated with a matrix of porous silicon, having an air-pore pitch of 1.58 μm, infiltrated with the nematic LC E7. The photonic band-gap transmission of this structure then becomes quite sensitive to temperature variations, through their effect on the effective refractive index of the LC as was demonstrated by the Japanese as well as the Canadian–German groups which studied this effect [208, 209].

C2.4.4.3 Bubble cross-connect arrays

A modified technology, originally developed for inkjet printers, to be used in an all-optical switch [210], is based on the formation of bubbles of gas by electrical heating, which then propel droplets of ink toward the printer paper. This printer technology is now quite reliable following two decades of development and perfection. The device generates bubbles in the same way, but uses them instead as a quickly appearing and disappearing gas–fluid optical interface.

At each switching point, two silica-based single-mode wave guides intersect at a fluid-filled trench such that the angle between each wave guide and the normal to the trench is greater than the angle at which total internal reflection begins for a gas-to-fluid interface. When a bubble is created at the intersection, light reflects off it; when the bubble disappears, light passes straight through. Switching time is of the order of 1 ms; the cross-talk was measured to be -70 dB. The researchers have fabricated an array of such elements into a 32×32 optical switch. The device is claimed to achieve up to a 20 year lifetime.

From the standpoint of SLM technology, we have a 2D array of binary-modulator switches capable of routing an incoming (1D) array of beams into an output array with a switching speed of 1 ms. This is a particular example of a ‘cross-bar switch’ implemented in a planar configuration. Similar functionality has been offered by using LC and micro-mirror (MEMS) technologies for optical communication (see section C2.4.5).

C2.4.4.4 Bio-chemical SLMs

This novel class of SLM technology is based on the spatial optical modulation induced as a result of chemical or bio-chemical spatial variations, which are usually converted to spatial modulation of the electric field/potential.

An earlier attempt to use bio-chemical agents to induce spatial optical modulation was made by using bacterio-rhodopsin molecules [211, 212]. However, while having scientific curiosity, this type of modulator has not yet been practically implemented.

Other more recent attempts have focused on the use of bio-chemical material used as bio- or chemical sensors to affect the optical property of the underlying optical modulator either directly or by spatially modulating an electric field or potential. A fibre-optic DNA sensor for nucleic acid determination was demonstrated using a fluorescent DNA stain [213]. Such a bio-chemo-optical sensor can be replicated in an array allowing the parallel detection of multiple forms of nucleic acids, and hence can be considered as a candidate technology for the ‘bio-chemical SLM’. Another such candidate based on an evanescent wave in a 1024 fibre optic array for the detection of oligo-nucleotides was demonstrated in 1996 [214]. Demonstrations of actual operation of arrays of bio-chemical optical modulators started to appear in the late 1990s. A fibre-optic-based micro-optode (i.e. optical micro-electrode) for the measurement of oxygen distribution was demonstrated in 1997 [215]. The researchers used a phase-modulation technique to determine the phase shift produced by the fluorescent stain, in response to a sinusoidal optical excitation. A biological sensing technique for the detection of specific bacteria was demonstrated using a micro-mechanical array of silicon nitride cantilever beams to which a specific immobilizing antibody of a certain bacterium (*E. coli*) was attached. The detection mechanism is based on the shift of the resonant frequency of the cantilevers which is measured by the optical beam. A limiting detection of 16 *E. coli* cells was demonstrated [216]. The simultaneous detection of six bio-hazardous agents using a cocktail of fluorescent antibodies coupled with the analysis of fluorescence intensity was demonstrated [217].

A surface-plasmon-based resonance was used to construct a bio-sensor array based on multiple analytes. By performing parallel, multiple spectroscopic analysis of the plasmon resonance for the various analytes, researchers demonstrated the bio-sensing ability to detect the affinity of the peptide sequence YGGFL to human β-endorphin, as well as the affinities between other types of bio-agent. The 1D array technology allows the simultaneous evaluation of up to 160 samples [218].

An inexpensive technology for the production of a colorimetric resonant optical bio-sensor, based on the use of multiple micro-titer wells and an array of fibre optic transceivers, was demonstrated [219]. The demonstration using 96 micro-titer plates showed the detection of a protein–protein affinity with an antibody detection sensitivity of 8.3 nM.

Finally, the use of a LC-based optical shutter array coupled to an aligned array of bio-sensing elements constructed on a planar wave guide was demonstrated [220]. The LC shutter array selectively controls the transmission of the fluorescence generated by the biosensor array, activated via the optical wave guide.

C2.4.5 SLM applications

C2.4.5.1 Optical communication, signal processing and interconnects

Overview

The most dramatic recent technology development relevant to the SLM field, since the last comprehensive report on SLM technology some 8 years ago [221], has been the explosive interest and use of communications. The most relevant applications of the SLM technology for this field are two fold:

(1) OXC switch arrays.

(2) SLM-based adaptive optical aberration correction systems for point-to-point optical communication (a part of the so-called ‘last mile communication’ problem).

A third application that is currently under development and which may be quite important for optical communication (OC) is that of adaptive ultrashort (femtosecond) pulse shaping. The latter is also an example for the use of optical signal processing (OSP) in optical communication. Indeed, since the communication field is by far the most dominant application for OSP and OI technologies, we will include a brief description of other OSP and OI applications as well within this section on optical communication.

Optical interconnect and cross-connect

In a typical operation, the optical cross-connect (OXC) unit enables a programmable interconnection between an array of *N* input optical channels (possibly implemented by *N* input fibre-optic channels) to an *N*-output channel array. It is desirable that the cross-connect switch allows spatial *and* wavelength-dependent routing operation. The role of an OXC unit for communication is to route each of the optical signals spatially to its designated end location, based on the signal (header) information and the particular operating wavelength of the channel. An obvious candidate of SLM technology to carry out this dynamic cross-connect function is LC technology. An earlier work used a PA LC-SLM for a 256-channel system [222]. In order to improve the performance of the slow-response NLC devices, an eight-channel device was demonstrated using FLC in 1995 [223].

A significant pioneering work in the area of LCDs for OI was done using both nematic [224, 225] and FLCDs [226], as well as a demonstration of polarization-insensitive switches [227], thereby overcoming one of the major weaknesses of the LC technology.

A study on the use of sub-wavelength LC-based diffractive optical elements for OXC applications was recently published [228].

The MEMS technology with its DMD representative detailed in section C2.4.3 earlier is another potential candidate for the OXC technology [229, 230], with numerous companies currently attempting to develop MEMS-based cross-connect switches. The relatively slow response coupled with the reflective-mode constraints is the main drawback of this technology.

The bubble-SLM technology mentioned in section C2.4.4 earlier is being developed for cross-connect application. A 32×32 switch has been demonstrated with a 10 ms response time [210, 231, 232].

The electro-hologram technology mentioned in section C2.4.4 earlier is under development for optical cross-connect applications. The main advantages of this technology over MEMS, LC, or the bubble technique is its relatively fast response (of the order of nanoseconds) as well as its unique wavelength selectivity.

An 8×8 optical crossbar is under development [233]. The OXC will consist of 64 inputs that can be routed to any one of the 64 outputs, using a hybrid III–V semi-conductor and Si CMOS flip-chip bonded technology. The optical switch will have 4096 detectors configured as a 64×64 array, each operating at $>200\,\mathrm{Mbit\ s^{-1}}$. Another effort has developed a 16×16 CMOS/GaAs OI operating at 0.83 μm [234].

Finally, there have been numerous demonstrations of VCSEL-based smart-pixel OI systems [235–238]. A free space, 64×64 optical cross-connect switch based on FLC-SLM was recently demonstrated [239]. The cross-connect interface can serve as a programmable parallel array processor for a variety of data processing applications.

Femtosecond pulse shaping

An important as well as an interesting application of the SLM technology for optical communication is the shaping of ultra-short, femtosecond optical pulses first suggested by Weiner [240] using a fixed mask, and later refined to include an LC-programmable mask or an SLM [241]. The concept is to spatially disperse the optical pulse using an appropriate grating and then spatially phase-modulate the dispersed signal pattern. The re-collection of the spatially phase-modulated pulse using a second grating, leads to a temporal re-shaping of the optical pulse (in particular, pulse compression), which can then be used for optical communication. In more recent developments, temporal as well as spatial pulse shaping was accomplished using SLMs [242, 243] with a recent demonstration of a 5 fs pulse compression [244].

C2.4.5.2 Display applications

This is obviously a very wide application field, with a huge billions of dollars market, encompassing numerous technologies. As we are interested only in reviewing the field from the SLM technology standpoint, we will confine ourselves to nonemissive or passive display technologies, omitting technologies such as LEDs/OLEDs, electro-luminescent and plasma displays. The discussion will therefore be limited to innovations in the area of LC- and MEMS-based displays.

One of the very recent applications of the PA LCLV discussed in section C2.4.3 earlier is that of electronic cinema [245]. Here, only high-brightness, high-resolution projection systems are required. A projection system based on combined LCLV technologies can provide up to 3000 ANSI lumens at a resolution of 2048 × 1536 pixels. A DMD-based projection system is in close competition.

Another important display technology for large-screen, video projection displays and for the electronic cinema application is the DMD, MEMS device discussed earlier [246]. The use of SLM panels (whether LCDs or digital mirrors) is implemented in a three-panel configuration, where either red, green and blue panels or some form of colour-subtraction configuration is used. This allows high brightness to be accomplished as the full bandwidth of the (white) light source is utilized simultaneously for all three channels. The disadvantage is in the complexity of the colour-separation system and the cost of three panels and their drivers. The other more compact and cost-effective method is to use a 'colour-sequential' scheme, where one broad-band SLM panel (either LCD or DMD) is used in conjunction with a colour-sequencing scheme where the red, green and blue portions of the filtered broad-band source are used sequentially to read out the panel. This requires a minimum of 180 Hz operation, which, although quite demanding, has been successfully demonstrated on LCD panels [247].

Finally, we should also mention the fast-growing field of head-mounted displays (HMDs) in which miniature LCD panels play an increasingly growing role [248–250]. LC-based, 3D stereoscopic displays are a derivative of both LCD and HMD technologies [251, 252] where the stereoscopic or, 3D depth visualization can be accomplished, e.g. by ascribing orthogonal polarizations to each of the ocular views using appropriate LC shutters [251]. Auto-stereoscopic systems where the stereoscopic imagery is embedded in the displayed imagery, obviating the need for the shutter glasses, has been a major thrust recently with DMD-type SLMs serving as the image source [252].

C2.4.5.3 Optical data processing

As pointed out in the previous section, novel methods and technologies related to the application field of OSP are covered within the former section on optical communication. This section covers innovations and updates related to the optical processing of spatial data information (the latter may either be optical or electronic) with an emphasis on image processing applications. One area, albeit not new, has been continuously attracting attention as one of the most efficient optical methods of image computation, namely the optical correlation technique. With the introduction of ultra-fast SLM technologies such as the MQW-SLM, a natural progression has been made to utilize this fast SLM in demonstrating high-throughput image correlation.

The group [148, 149] claims a correlation throughput of up to 600 000 correlations s^{-1} with 128 × 128 images. This translates into an equivalent computational throughput of: $P_{\mathrm{comp}} = 2N^2 \log N \cdot f \sim 115\,\mathrm{GOPS}$, where $N = 128$; $f = 600\,000$, and GOPS = giga binary operations s^{-1}.

While this obviously is a very respectable computational throughput, it is uncomfortably close (to within a factor of approximately ×10) to throughputs attainable by current dedicated ULSI, system-on-chip (SOC) processors. This state of affairs represents the on-going dilemma of optical processing—the very fast moving electronic competition.

C2.4.5.4 Adaptive and programmable optics applications

Here, we focus on the use of dynamic, adaptive optics and programmable optic techniques for noncommunication applications including astronomical and vision systems.

Adaptive optical correction

The use of adaptive optics for atmospheric aberration corrections has been intensively researched in the last four decades or so. The particular use of SLM technology as a key tool in adaptive, real-time correction systems was given in detail by Pepper *et al* [221, chapter 14].

Phase conjugation, whereby a phase-aberrated wave-front is 'cleaned up', is an attractive way of correcting phase aberrations of optical beams. There have been several attempts to use the PA LCLV technology, through its capability of real-time phase conjugation, to demonstrate such an action first by using nematic LCs [253, 254] and later by using FLC modulators [255]. More recent use of SLM technology for real-time aberration correction appears in references [256, 257].

A recent application of this technology has emerged in the area of visual system examination and correction. A retinal imaging system, which uses a deformable mirror-type SLM for the correction of the eye-lens aberration was recently developed [258]. Here, a Hartmann–Shack-type wave-front sensor senses the ocular lens aberrations. These wave-front distortion data are subsequently fed to a deformable mirror SLM, to correct adaptively the retinal image. This system allows a clear, nonaberrated imaging of the retina for performing retinal examination and subsequent medical treatment.

The next stage in the use of this technology is an adaptive visual correction system, in which this real-time SLM-based, adaptive optics technique will be used to attain a 'super-normal vision', thus allowing a near-diffracted-limited vision to be achieved.

Programmable optical elements

This application, which is related to the previous 'adaptive optics' one, is based on the use of phase-modulating SLMs to perform programmable optical element functions e.g. variable focal-length lenses or variable slope prisms used for beam-steering. Such devices often make use of the fact that a given optical phase profile required for a desired shaping of a monochromatic optical beam wave-front can be substituted by its 'modulo-2π' representation. Thus for example, a linear phase profile, $\phi(x)$, of a glass prism is described by: $\phi(x) = \tan(\alpha)xn_0$, where α is the apex angle of the prism, n_0 is the glass refractive index, and x represents a point along the length of the prism, will have the same effect on a monochromatic beam at λ_0 as a phase profile which 'resets' to: $\phi(x) = 0$, at x-locations in which $\phi(x)$ satisfies: $\phi(x)/\lambda_0 = 2\pi n$, where $n = 1, 2, \cdots$. Note that this allows relatively thin optical elements with thickness of the order of λ, namely a few micrometres, to be substituted for the traditional thick (mm–cm) glass elements. Also note, however, that such substitution is only valid for the particular wavelength λ_0 for which $\phi(x)/\lambda_0 = 2\pi n$. These optical elements are commonly referred to as 'binary optics' or 'kino-form optical elements' [259, 260].

The ability to structure wavelength-thin, optical elements opened the way for using phase modulating SLMs thereby allowing *programmable* optical elements to be developed.

Applications of such dynamic kino-form optical elements include LC beam-steering devices [261, 262], as well as variable focal-length lenses [263–265]. The main drawback of these dynamic programmable elements, as in the static kino-form elements, is their limitation in *monochromatic* beam shaping applications.

C2.4.5.5 Wavelength image converters

An additional important application of SLMs is the conversion of an imagery from one wavelength to another. This can be done using either a single, photo-activated SLM whose photo-substrate and optical modulators operate at different wavelengths. Alternatively, one can use an imager device to record the imagery at a certain wavelength and relay the electronic signal to an electrically addressed SLM operating at a desired output wavelength.

A commonly used example of the latter is the regular TV monitor which relays the image picked up by a video camera and displays the colour version video on the screen. Although in this case both the input and the output imagery are formally based on red–green–blue (RGB) visible channels, we can easily manipulate the TV colour range, and thus can effectively perform an image wavelength conversion operation with this system.

PA LCLVs of the type discussed in section C2.4.3 earlier have been successfully used in converting visible imagery to IR video scenes for IR scene simulation applications [266]. Such devices were also used for visible- to near-IR image conversion [267].

The high spatial resolution and array sizes attainable in some SLM technologies, combined with the capability of UV modulation, naturally calls for the use of this technology to perform programmable, real-time photo-lithography [268]. The use of an NLC-SLM with 600×800 pixel resolution, 27 mm × 20 mm aperture, over 100:1 contrast and 7/25 ms rise/decay response times at the argon ion wavelength of 351.1 nm, were used to form a 3D hologram in a photopolymer material. The hologram was subsequently read using the 633 nm of a He–Ne laser [269].

C2.4.6 Fundamental limits and future trends

C2.4.6.1 Performance trade-offs and fundamental limits

Some important interplays or trade-offs exist among the main SLM performance parameters. It is important to understand those limits in order to realistically design an SLM or make a sound projection of its expected performance. The main parameter trade-offs of SLMs are briefly described below.

Dynamic range/speed

This is perhaps the most significant trade-off in the SLM technology. It is somewhat intuitive in that a large electro-optic coefficient, which is related to the dynamic range attainable by the optical modulator system, will be associated with a reduced speed of response, and vice versa. A good example for two modulator systems showing this trade-off would be, on the one hand, the class of nematic liquid crystal material featuring an enormously high (second-order) electro-optic effect with an effective half-wave voltage of around 2 V, but with a typically low frequency response of around 100 Hz. (The half-wave voltage can roughly be taken as inversely proportional to the electro-optic coefficient.) This is contrasted with solid-state electro-optic crystals such as the KD^*P or $LiNbO_3$ capable of responding to bias frequencies well above 1 MHz, but with the penalty of a very large, 1000 V half-wave voltage. To gain some insight into this trade-off—reminiscent of the well-known gain–bandwidth product (GBWP) in electronic devices—it can be argued that under certain hypotheses this trade-off may in fact, obey a ‘universal’ form of the GBWP behaviour, for a variety of optical modulator material systems, where the gain is actually represented by the generalized susceptibility relevant to the particular modulator material system (e.g. the electro-optic coefficient, for solid state EO modulators), and where the bandwidth is the total frequency bandwidth of the modulator system, comprised of both the optical (spectral) bandwidth of the modulator *and* the electric field, RF-modulating frequency (see appendix).

Resolution–speed trade-off

This constraint is originated in the dynamic range–speed trade-off, as discussed earlier. Thus, for very fast-responding materials (e.g. solid-state electro-optic crystals) the high biasing voltage required, due to their low EO coefficients, as a result of the above, dynamic range–speed trade-off, will necessitate the use of large pixel sizes to avoid field break-downs.

Dynamic range–resolution

A common manifestation of this trade-off is the familiar 'blooming' effect, which occurs in the 'over-activated' regions of the SLM pixels. This is usually due to the finite charge-holding capacity of the driver array (both in PA and electrically addressed devices). The excessive pixel driver charge, resulting from over-driving the pixel in attempting to attain high output signal (or equivalently, high dynamic range), spills over to adjacent pixels, resulting in the 'blooming effect'. However, this trade-off can also be 'utilized' in a constructive manner by using a cluster of pixels in a high-resolution, binary-mode device to produce effectively a grey-level modulation at the lower resolution defined by the pixel cluster, in a similar way to half-tone image techniques.

Sensitivity–speed

This trade-off (for PA devices) is usually the consequence of the constancy of the energy flux required for the photoactivations ($\mathrm{J\,cm^{-2}}$), rather than that of the power flux ($\mathrm{W\,cm^{-2}}$). Examples are PA SLMs, in which we must compensate for low input power flux levels, by reducing their frame rate or integration time. Another well-known example is the reciprocal relationship between the response time for the formation of a grating in a photo-refractive material, and the power level of the input signal. In general, the consequence is the necessity to use longer integration periods to accumulate sufficient signal charge and thus to lower the speed or the frame rate of the device. This is similar to the limitation of imagers under low-level illumination.

C2.4.6.2 Future trends in SLM technology

The last decade has witnessed tremendous changes in three main optics-related areas, namely, communications, information processing and display technologies. The explosive growth and the prospects of even more dramatic future developments are certainly expected to impact the trend in the related SLM technology. In addition, the trend of developing artificial man-made optical materials, which started around two decades ago, has led to a very successful development of, in fact, the fastest optical modulator in existence today, namely, the MQW modulator. Therefore, a continuation of all these trends in the next decade is predicted. The development of OXC-SLM technology, which has already seen the development and adaptation of MEMS-based devices, will continue to expand with a thrust towards the development of novel SLM technologies such as the Bubble Array or the electro-hologram devices. We have already witnessed the redirection of the historic optical data processing technology into the use of OIs. It is predicted that this will continue as a thrust for the development of SLMs in information processing systems. The development of relatively slow ($\sim$1 ms) phase-modulating devices for adaptive and programmable optical element applications will continue to grow. At the same time, it is expected that the effort in developing faster modulators will be expanded. In the booming and attractive field of displays, active arrays such as organic LEDs (OLEDs) may well challenge the traditional passive, SLM-type devices such as LCDs. Finally, with the strong interest in bio-chemical data processing and biological sensors—enhanced by the explosion in genetic research as well as the recent threats of weapons of mass destruction, one can see a significant thrust in the area of bio-chemical SLMs (see section C2.4.4.4) to allow a fast, parallel processing and analysis of bio-chemical data.

References

[1] Efron U (ed) 1990 *Spatial Light Modulators-III Proc. SPIE vol 1150*, (Bellingham, WA: SPIE)
[2] Optical Society of America (OSA) 1990 *Topical Meeting on Spatial Light Modulators* Sept. 1990, Incline Village, NV
[3] Optical Society of America (OSA) 1993 *Topical Meeting on Spatial Light Modulators* March 1993, Palm Springs, CA
[4] Efron U (ed) 1995 *Spatial Light Modulator Technology* (New York: Dekker)
[5] Roy A and Singh K 1995 Spatial light modulators and their applications:A bibliographical revier for the years 1990–1991 *Atti dela Fondazione Giorgio Ronchi* **51** 529–601
[6] Burdge G (Chairman) 1997 Optical Society of America (OSA), *Topical Meeting on Spatial Light Modulators*
[7] *Applied Optics: Special Issue on Spatial Light Modulators*, *vol 37* (November 1998)
[8] Sutherland R L (ed) 1998 *Spatial Light Modulators* (*Proc. SPIE vol 3292*)
[9] *Diffractive and Holographic Technologies, Systems, and Spatial Light Modulators* IV 1999 (*Proc. SPIE vol 3633*)
[10] *Diffractive/Holographic Technologies and Spatial Light Modulators* VII 2000 (*Proc. SPIE vol 3951*)
[11] Efron U (ed) 2001 *Spatial Light Modulators: Technology and Applications*, *Proc. SPIE*
[12] Sayyah K and Efron U 1996 Optically addressed spatial light modulator with high photosensitivity and intensity adaptation range *Opt. Lett.* **21** 1384–1386
[13] Esener S C, Wang J H, Drabik T J, Title M A and Lee S H 1987 One-dimensional silicon PLZT spatial light modulator *Opt. Eng.* **26** 406–413
[14] Donjon J, Dumont F, Grenot M, Hazan J P, Marie G and Pergrale J 1973 *IEEE Trans. Electron Devices* **20** 1037
[15] Casasent D 1978 *Opt. Eng.* **17** 344
[16] Warde C, Weiss A M, Fisher A D and Thackara J I 1981 *Appl. Opt.* **20** 2066–2074
[17] Schwartz A, Yang X Y and Warde C 1984 Electron beam addressed, microchannel spatial light modulator *Proc. SPIE* **465** 23–28
[18] Moddel G 1995 Ferroelectric liquid crystal spatial light modulators *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker) pp 287–359
[19] Chigrinov V G 1999 *Liquid Crystal Devices* (Boston: Artech House)
[20] Huang X Y, Miller N, Khan A, Davis D, Doane J W and Yang D K 1998 Grey-scale of bistable reflective cholesteric displays *SID '98 Digest* 810–813
[21] Hornbeck L J 1990 Deformable mirror spatial light modulator *Spatial Light Modulators and Applications III (Proc. SPIE Vol. 1150)*, ed U Efron pp 86–103
[22] Van Kessel P F and Hornbeck L J 1998 A MEMS-based projection system *Proc. IEEE* **86** 1686–1704
[23] Bloom D M 1997 Grating light valve: revolutionizing display technology *Proc. SPIE* **3013** 165–171
[24] Kubota S R 2002 The grating light valve projector *Opt. Photonics News* **13** 50–60
[25] Fisher A D, Ling L C, Lee J N and Fukuda R C 1986 Photoemitter membrane light modulator *Opt. Eng.* **25** 261–268
[26] Hornbeck L J 1983 *IEEE Trans Electron Devices* **30** 539
[27] Pape D R 1984 An optically addressed membrane spatial light modulator *Proc. SPIE* **465** 17–22
[28] Kingston R H, Burke B E, Nichols K B and Leonberger F J 1982 Spatial light modulation using electro-absorption in GaAs CCD *Appl. Phys. Lett.* **41** 413–415
[29] Kingston R H, Burke B E, Nichols K B and Leonberger F J 1984 An electro-absorptive CCD spatial light modulator *Proc. SPIE* **465** 9–11
[30] Miller D A B 1987 Quantum wells for optical information processing *Opt. Eng.* **26** 368
[31] Mast F and Waser R 1980 Optischer bildverstärker *European Patent Specification 0029-006* (Gretag AG)
[32] Hess K, Dändliker R and Thalman R 1987 Deformable surface spatial light modulator *Opt. Eng.* **26** 418–422
[33] Lea M C 1983 *Appl. Phys. Lett.* **43** 738
[34] Lea M C 1984 Optical modulators based on electro-capillarity *Proc. SPIE* **465** 12–16
[35] Horwitz B A and Corbett F J 1978 *Opt. Eng.* **17** 353
[36] Petrov M P 1980 Diffractive and dynamic properties of photosensitive electro-optic media *Am. Inst. Phys. Conf. Proc.* **65** 493–507
[37] Casasent D 1981 Soviet PRIZ spatial light modulator *Appl. Opt.* **20** 3090–3092
[38] Petrov M P 1984 Physical basis of operation of the PRIZ spatial light modulator *Optik* **67** 247–256
[39] Marrakchi A, Tanguay A R, Yu J and Psaltis D 1985 *Opt. Eng.* **24** 124
[40] Psaltis D, Yu J, Marrakchi A and Tanguay A R 1984 Photorefractive incoherent to coherent optical conversion *Proc. SPIE* **465** 2–8
[41] Rabinovich W S, Bowman S R, Katzer D S and Kyono C S 1995 Intrinsic MQW spatial light modulators *Appl. Phys. Lett.* **66** 1044–1046

[42] Lahiri I, Kwolek K M, Nolte D D and Melloch M R 1995 Photorefractive p–i–n diode MQW spatial light modulator *Appl. Phys. Lett.* **67** 1408–1410

[43] Pape D R 1995 Acousto optic Bragg cell devices *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker) pp 415–442

[44] 1989 MQW optical modulator structures using surface acoustic wave-induced Stark effect *IEEE Photon. Technol. Lett.* **1** 307–309

[45] Davis J A and Waas J M 1990 Current status of the magneto-optic spatial light modulator *Proc. SPIE* **1150** 27–45

[46] Strome D H 1984 Cinematic infrared scene simulator based on vanadium dioxide spatial modulator *Proc. SPIE* **465** 192–196

[47] Sasaki A *et al* 1980 *Proc. Soc. Inf. Display* **21** 341

[48] Wallace J 2000 *Laser Focus World* **36**(5)

[49] Xu H, Davey A B, Wilkinson T D and Crossland W A 1999 Plasmon effect: optically enhancing the small electro-optical effect of a fast-switching liquid crystal mixture *Opt. Eng.* **39** 1568–1572

[50] Leonard S W, Mondia J P, VanDriel H M, Toader O, John S, Busch K, Birner A, Gösele U and Lehman V 2000 Tunable two-dimensional photonic crystals using liquid crystal infiltration *Phys. Rev.* **B61** R2389–R2392

[51] Friend R H 1998 Organic electroluminescent displays *Soc. Inf. Display Lecture Notes* **2** 1–27

[52] Burroughes J H *et al* 1990 Light emitting diodes based on conjugated polymers *Nature* **347** 539–541

[53] Iga K, Koyama F and Kinoshita S 1988 VCSEL array surface emitting semiconductor lasers *IEEE J. Quantum Electron.* **24** 1845–1855

[54] Chang-Hasnain C J *et al* 1991 Multiple wavelength tunable, surface emitting laser arrays *IEEE J. Quantum Electron.* **27** 1368–1376

[55] Efron U 1991 Liquid crystals materials devices and applications *Handbook of Microwave and Optical Components* vol 4, ed K Chang (New York: Wiley) pp 372–374

[56] Efron U 1991 Liquid crystals materials devices and applications *Handbook of Microwave and Optical Components* vol 4, ed K Chang (New York: Wiley) pp 370–371

[57] Grinberg *et al* 1975 *Opt. Eng.* **14** 217

[58] Chigrinov V G 1999 *Liquid Crystal Devices* (Boston: Artech House) pp 234–240

[59] Clark N A and Lagerwall T 1980 *Appl. Phys. Lett.* **36** 899

[60] Fergason J L 1985 *Soc. Inf. Display Digest* **16** 68

[61] Vaz N A 1989 *Proc. SPIE* **1080** 2

[62] Ichikawa H, Kataoka H, Oka T, Iida M, Fujioka T and Ino M 2000 Low power poly-Si reflective colour AMLCD with 1024 × 480 pixels *Soc. Inf. Display Digest* 1203–1207

[63] Mishima Y, Nakayama T, Suzuki N, Ohta M, Endoh S, Iwakabe Y and Kagawa H 2000 Development of a 19-in.-diagonal UXGA super TFT-LCM applied with super-IPS technology *Soc. Inf. Display Digest* **19** 260–265

[64] Schadt M, Scmitt K, Kozinkov V and Chigrinov V G 1992 Photo-polymer alignment methods *Japan. J. Appl. Phys.* **31** 2155

[65] de Gennes P G and Prost J 1993 *The Physics of Liquid Crystals* (Oxford: Oxford University Press)

[66] Blinov L M 1983 *Electro-Optical and Magneto-Optical Properties of Liquid Crystals* (New York: Wiley)

[67] Chandrasekhar S 1992 *Liquid Crystals* 2nd Ed (Cambridge: Cambridge University Press)

[68] Efron U 1991 Liquid crystals materials devices and applications *Handbook of Microwave and Optical Components* vol 4, ed K Chang (New York: Wiley)

[69] Chigrinov V G 1999 *Liquid Crystal Devices* (Boston: Artech House)

[70] Khoo I C and Wu S T 1993 *Optics and Non-Linear Optics of Liquid Crystals* (Singapore: World Scientific)

[71] Chigrinov V G 1999 *Liquid Crystal Devices* (Boston: Artech House) pp 100–134

[72] Janning J L 1972 Thin film surface orientation for liquid crystals *Appl. Phys. Lett.* **21** 173

[73] Becker M E *et al* 1986 Alignment properties of rubbed polymer surfaces *Mol. Cryst. Liq. Cryst.* **132** 167–180

[74] Schadt M *et al* 1992 Surface induced parallel-alignment of liquid crystals by linearly polarized photopolymers *Japan. J. Appl. Phys.* **31** 2155–2164

[75] Kataoka S, Taguchi Y, Iimura Y, Kobayashi S, Hasebe H and Takatsu H 1997 *Mol. Cryst. Liq. Cryst.* **292** 333

[76] Ikeno Y *et al* 1988 Electrooptic bistability of a ferroelectric liquid crystal device prepared using polyimide Langmuir-Blodgett orientation films *Japan. J. Appl. Phys.* **27** L475

[77] Lien S-C A *et al* 1998 Active-matrix display using ion-beam-processed polyimide film for liquid crystal alignment *IBM J. Res. Dev.* **42**(3)

[78] Nakamura M and Ura M 1981 Alignment of nematic liquid crystals on ruled grating surfaces *J. Appl. Phys.* **52** 210

[79] Uchida T *et al* 1980 *Japan. J. Appl. Phys.* **19** 2127

[80] Seki H *et al* 1990 Tilted-homeotropic alignment of liquid crystal molecules using the rubbing method *Japan. J. Appl. Phys.* **29** L2236–L2238

[81] Lackner A M *et al* 1990 *Dig. Soc. Inf. Display SID 90* vol XXI p 98

[82] Bleha W P 2000 D-ILA technology for electronic cinema *Dig. Soc. Inf. Display SID 2000* vol XXXI pp 310–313

[83] Miller L *et al* 1991 Method for tilted alignment of liquid crystals with improved photostability, US Patent specification 5 011 267

[84] Lu M *et al* 2000 Homeotropic alignment by single oblique evaporation of SiO_2 and its application to high resolution microdisplays *Dig. Soc. Inf. Display SID 2000* vol XXXI pp 446–449

[85] Yoshida H and Koike K 1997 *Japan. J. Appl. Phys.* **36** L428–L431

[86] Furumi S *et al* 1999 *Appl. Phys. Lett.* **74** 2438–2440

[87] Park B *et al* 1999 *J. Appl. Phys.* **86** 1854–1859

[88] Kimura M 2000 New photo alignment technology based on (4-chalconyloxy) alkyl groups *Dig. Soc. Inf. Display SID 2000* vol XXXI pp 438–441

[89] Gooch C H and Tarry H A 1975 *J. Phys.* **D8** 1575

[90] Efron U and Wu S T 1985 *Opt. Eng.* **24** 111

[91] Konforti N, Marom E and Wu S T 1988 Phase only modulation with twisted nematic liquid crystal–spatial light modulators *Opt. Lett.* **13** 251–253

[92] Blinov L M 1983 *Electro-Optical and Magneto-Optical Properties of Liquid Crystals* (New York: Wiley) chapter 6

[93] de Vries H 1951 *Acta Crystallogr.* **4** 219

[94] Dierking I *et al* 1996 *J. Appl. Phys.* **81** 3007–3014

[95] Yang D-K *et al* 1994 *J. Appl. Phys.* **76** 1331–1333

[96] Clark N A and Lagerwall S T 1980 *Appl. Phys. Lett.* **36** 899

[97] Displaytech FLC Devices

[98] Garoff S and Meyer B 1979 Electroclinic effect at the A–C phase change in a chiral smectic liquid crystal *Phys. Rev.* **A19** 338–347

[99] Funfschilling and Schadt 1989 *J. Appl. Phys.* **66** 3877–3882

[100] Fergason J L 1985 *SID Int. Symp. Dig. Tech. Papers* **16** 68

[101] Doane J W, West J L, Golemme A, Whitehead J B Jr and Wu B-G 1988 *Mol. Cryst. Liq. Cryst.* **165** 511

[102] Crawford G P 1996 Liquid crystal polymer dispersions for reflective flat-panel displays *Soc. Inf. Display, Tech. Digest* F/4-1–F/4-47

[103] Kitzerov H S 1994 Polymer dispersed liquid crystals, from the nematic curvilinear phase to ferroelectric films *Liq. Cryst.* **16** 1–31

[104] Coates D 1995 Polymer dispersed liquid crystals *J. Mater. Chem.* **5** 2063–2072

[105] Xu M and Yang D 1999 PSCT *Proc. SID* 950–953.

[106] Soc. Inf. Display, *Int. Symp. Digest* 1996; Soc. Inf. Display, *Int. Symp. Digest* 1997; Soc. Inf. Display, *Int. Symp. Digest* 1998; Soc. Inf. Display, *Int. Symp. Digest* 1999; Soc. Inf. Display, *Int. Symp. Digest* 2000

[107] Grinberg J *et al* 1975 *Opt. Eng.* **14** 217

[108] Displaytech Co. (Longmont, CO, USA), LightView –QVGA, *Display Module, Model QDM-0076-MV5* www.Displaytech.com

[109] Kopin Co. (Taunton, MA, USA) *CyberDisplay-1280-Mono* www.kopin.com

[110] Akimoto O and Hashimoto S 2000 A 0.9-in UXGA/HDTV FLC micro-display *Soc. Inf. Display, Tech. Digest* pp 194–197

[111] Date M, Hisaki T, Naito N, Nakadaira A, Suyama S, Tanaka H, Uehira K and Koshiishi Y 2000 Direct-viewing display using alignment-controlled PDLC and holographic PDLC *Soc. Inf. Display Tech. Digest* pp 1184–1188

[112] Efron U, Sayyah K, Byles W R, Goodwin N W, Forber R A, Wu C S and Welkowsky M S 1991 The CCD-addressed liquid crystal light valve—an update *Proc. SPIE* **1455** 237–247

[113] Sterling R D and Bleha W P 2000 D-ILA™ technology for electronic cinema *Soc. Inf. Display Digest* pp 310– 313

[114] Sayyah K, Efron U and Forber R A 1995 Color-sequential crystalline Si-LCLV-based projector for consumer HDTV *Proc. Soc. Information Displays (SID) Digest* p 520

[115] Kanichi J (ed) 1991 Amorphous Si for optically-addressed spatial light modulators *Amorphous and Microcrystalline Semiconductor Devices: Optoelectronic Devices* (Norwood, MA: Artech House) pp 369–412

[116] Aubourg P, Huignard J P, Hareng M and Mullen R A 1982 Liquid crystal light valve using bulk, monocrystalline $Bi_{12}SiO_{20}$ *Appl. Opt.* **21** 3706–3712

[117] Moddel G 1995 Ferroelectric LC spatial light modulators *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker) pp 310–311

[118] Hornbeck L J 1983 *IEEE Trans. Electron. Devices* **30** 539

[119] Brooks R E 1984 Micro-mechanical light modulators for data transfer and processing *Proc. SPIE* **465** 46–54

[120] Hornbeck L J 1990 Deformable mirror spatial light modulator *Proc. SPIE* **1150** 86–102

[121] Van Kessel P F and Hornbeck L J 1998 A MEMS-based projection system *Proc. IEEE* **86** 1686–1704

[122] Bloom D M 1997 Grating light valve: revolutionizing display technology *Proc. SPIE* **3013** 165–171

[123] Kubota S R 2002 The grating light valve projector *Opt. Photonics News* **13** 50–60

[124] Hwang Kyu-Ho, Song Yong-Jin and Kim Sang-Gook 1998 Thin film micromirror array for high-brightness projection displays *Japan. J. Appl. Phys.* **37** 7074–7077

[125] Sakarya S, Vdovin G and Sarro P M 2002 Technology of reflective membranes for spatial light modulators *Sens. Actuators* **A97–98** 468–472

[126] Riesenberg R 2001 Micro-mechanical slit positioning system as transmissive SLM *Proc. SPIE* **4457** 197–204

[127] Chemla S *et al* 1983 *Appl. Phys. Lett.* **42** 864–865

[128] Miller A B *et al* 1983 *Appl. Phys. Lett.* **42** 925

[129] Kingston R H *et al* 1982 Spatial light modulator using electro-absorption in a GaAs CCD *Appl. Phys. Lett.* **41** 413–415
[130] Hsu T Y *et al* 1988 *Opt. Eng.* **27** 372–384
[131] Miller D A B *et al* 1985 *Phys. Rev.* **B32** 1043–1060
[132] Chang Y C, Schulman J N and Efron U 1987 *J. Appl. Phys.* **62** 4533
[133] Ploog K and Döhler G H 1983 Compositional and doping superlattices in III–V semiconductors *Adv. Phys.* **32** 285–359
[134] Döhler G H 1986 Light generation, modulation and amplification by nipi doping superlattices *Opt. Eng.* **25** 211–218
[135] Kiesel P *et al* 1993 High speed and high contrast electro-optic modulators based on nipi doping superlattices *Superlattices Micro-Structures* **13** 21–24
[136] Wu W-Y *et al* 1987 *Appl. Phys. Lett.* **51** 710–712
[137] Miller D A B *et al* 1986 *Appl. Phys. Lett.* **49** 821
[138] Efron U and Livescu G 1995 Multiple quantum well spatial light modulators *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker)
[139] Yan R H, Simes R J and Coldren L A 1989 Electroabsorptive Fabry–Perot reflection modulators with asymmetric mirrors *IEEE Photonic Technol. Lett.* **1** 273
[140] Lin *et al* 1994 *Appl. Phys. Lett.* **65** 1242
[141] Hsu T Y, Efron U and Wu W Y 1988 Amplitude and phase modulation in a 4-μm-thick GaAs/AlGaAs multiple quantum well modulator *Electron. Lett.* **24** 603–604
[142] Efron U and Livescu G 1995 Multiple quantum well spatial light modulators: materials devices and applications *Spatial Light Modulators and Applications*, ed U Efron (New York: Dekker) pp 243–247
[143] Hsu T Y and Efron U 1989 Review of multiple quantum well spatial light modulators *Proc. SPIE* **1150**
[144] Ahearn J S *et al* 2001 Multiple quantum well spatial light modulators for optical data processing and beam steering applications *Proc. SPIE* **4457** 43–53
[145] Livescu G *et al* 1988 *Opt. Lett.* **13** 297
[146] Hsu T Y and Efron U 1989 *Proc. SPIE* **1150** 80
[147] Worchesky T L *et al* 1996 *Appl. Opt.* **35** 1180–1186
[148] Trezza J A *et al* 1998 *Proc. SPIE* **3490** 78–81
[149] Kang K *et al* 1999 *Proc. SPIE* **3715** 97–107
[150] Krishnamoorthy A V *et al* 1999 *IEEE J. Sel. Top. Quantum Electron.* **5** 261–275
[151] Bowman S R *et al* 1998 *J. Opt. Soc. Am.* **B15** 640–647
[152] Walker A C *et al* 1999 *IEEE J. Sel. Top. Quantum Electron.* **5** 236–249
[153] Junique S *et al* 2001 MQW-SLM for optical information processing *Proc. SPIE* **4457** paper No 10
[154] Lenslet Laboratories-Ramat-Gan Israel: Private Communication 2002 www.lenslet.com.
[155] Efron U (ed) 1984 *Spatial Light Modulators and Applications* vol 465 (Bellingham, WA: SPIE) pp 2–9, 23–29, 82–97
[156] Efron U and Warde C (eds) 1986 Materials and devices for optical information processing: Special issue *Opt. Eng.* **25** 250–261
[157] Efron U (ed) 1987 *Spatial Light Modulators and Applications* vol 825 (Bellingham, WA: SPIE) pp 106–113; Efron U (ed) 1987 *Spatial Light Modulators and Applications* vol 825 (Bellingham, WA: SPIE) pp 88–94, 198–206
[158] Esener S 1995 Smart pixels: technology and applications to parallel computing *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker) pp 449–453
[159] Yariv A and Yeh P 1984 *Optical Waves in Crystals* (New York: Wiley–InterScience) p 231
[160] Esener S C, Wang J H, Drabik T J, Title M A and Lee S H 1987 One dimensional silicon-PLZT spatial light modulator *Opt. Eng.* **26** 406–413
[161] Debye P and Sears F W 1932 *Proc. Natl Acad. Sci. USA* **18** 409–414
[162] Saleh B E A 1991 *Fundamentals of Photonics* (New York: Wiley–InterScience) chapter 20
[163] Yariv A and Yeh P 1984 *Optical Waves in Crystals* (New York: Wiley–InterScience) chapter 8
[164] Pape D 1995 Acousto-optics Bragg-cell devices *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker)
[165] Pulliam G R, Ross W E, MacNeal B E and Bailey R F 1982 *J. Appl. Phys.* **53** 2754
[166] Davis J A and Waas J M 1990 Current status of the magneto-optic spatial light modulator *Proc. SPIE* **1150** 27–45
[167] Ross W E, Psaltis D and Anderson R H 1983 Two-dimensional magneto-optic spatial light modulator for signal processing *Opt. Eng.* **22** 485–490
[168] Park J, Cho J, Nishimura K and Inoue M 2002 Magneto-optic spatial light modulator for volumetric digital recording system *Japan. J. Appl. Phys.* **41** 1813–1816
[169] Park J, Cho J, Nishimura K and Inoue M 2002 New drive line shape for reflective magneto-optic spatial light modulator *Japan. J. Appl. Phys.* **41** 2548–2551
[170] Ashkin A *et al* 1966 *Appl. Phys. Lett.* **9** 72
[171] Glass A M *et al* 1972 *Natl Bur. Stand. Spec. Publ. 372* 15
[172] Guenter P 1982 *Phys. Rev.* **93** 199
[173] Wood G L *et al* 1995 Photorefractive materials *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker)
[174] Yeh P 1989 *Proc. SPIE* vol 825, ed U Efron pp 96–100

[175] For a more comprehensive review of the use of PR materials for phase conjugation see e.g. Feinberg J 1985 Optical phase conjugation in photo-refractive materials *Optical Phase Conjugation*, ed R A Fisher (New York: Academic)

[176] This technology area has been extensively published. For a recent article see e.g., Burr G W *et al* 2001 Volume holographic storage at an area density of 250 Gpixels/in^2 *Opt. Lett.* **26** 444–446

[177] Iemmi C and La Mela C 2002 Phase only photo-refractive joint transform correlator *Opt. Commun.* **209** 255–263

[178] Choiou A E and Yeh P 1986 *Opt. Lett.* **11** 461

[179] O'Meara T R, Pepper D M and White J O 1985 Applications of nonlinear, optical phase conjugation *Optical Phase Conjugation*, ed R A Fisher (New York: Academic)

[180] Yau H F, Lee H Y and Cheng N J 1999 *Appl. Phys.* **B68** 1055

[181] Delaye P and Roosen G 1999 Evaluation of a photo-refractive two-beam coupling novelty filter *Opt. Commun.* **165** 133–151

[182] Marrakchi A, Tanguay A R Jr, Wu J and Psaltis D 1984 *Proc. SPIE* vol 465, ed U Efron pp 82–96

[183] Ma J, Liu L, Wu S, Wang Z and Xu L 1989 *Opt. Lett.* **14** 572

[184] Sharp E J, Wood G L, Clark W W III, Salamo G J and Neurgaonkar R R 1992 *Opt. Lett.* **17** 207

[185] Bowman S R, Rabinovich W S, Beadie G, Kirkpatrick S M, Katzer D S, Ikossi-Anastasiou K and Adler C L 1998 *J. Opt. Soc. Am.* **B15** 640–647

[186] Esener S 1995 Smart pixels: technology and applications to parallel computing *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker)

[187] Krishnammorthy A *et al* 1997 Progress in optoelectronic VLSI smart pixel technology based on GaAs/GaAlAs MQW modulators *Int. J. Optoelectron.* **11** 181–198

[188] Lin T H *et al* 1990 *Appl. Opt.* **29** 1595

[189] Cotter L K, Drabnik T J, Dillon R J and Handschy M A 1990 *Opt. Lett.* **15** 291

[190] D'Asaro L A, Chirovsky L M F, Laskowski E J, Pei S S, Woodward T K, Lentine A L, Leibenguth R E, Focht M W, Freund J M, Guth G D and Smith L E 1993 Batch fabrication and operation of GaAs/AlGaAs field effect transistor self-electro-optic effect device (FET-SEED) smart pixel arrays *IEEE J. Quantum Electron.* **29** 670–675

[191] Neff J *et al* http://wwwocs.colorado.edu/research/fsoi/research6.html

[192] Leibowitz B, Boser B E and Pister K S J CMOS 'smart pixel' for free-space optical communication *Proc. SPIE—The International Society for Optical Engineering* vol 4306A (*Electronic Imaging* '01) (San Jose, CA, January 2001)

[193] Efron U, Sayyah K, Byles W R, Goodwin N W, Forber R A, Wu C S and Welkowsky M S 1991 The CCD-addressed liquid crystal light valve—an update *Proc. SPIE* **1455** 237–247

[194] Efron U, Davidov I, Sinelnikov V and Friesem A 2001 CMOS/LCOS-based image transceiver device *Proc. SPIE* **4457** 188–196

[195] Walker A C *et al* 1998 Opto-electronic systems based on InGaAs-complementary metal oxide semiconductor smart pixel arrays and free space optical interconnects *Appl. Opt.* **37** 2822–2830

[196] Krishnamoorthy A V *et al* 1997 Dual function detector–modulator smart pixel *Appl. Opt.* **37** 4866–4870

[197] Zhang L, Hong S, Min C, Alpasan Z Y and Sawchuk A A 2001 Optical multi-token-ring networking using smart pixels with field programmable gate arrays *Proc. SPIE* 4470

[198] Kane J S, Kincaid T G and Hemmer P 1998 Optical processing with feedback using smart pixel spatial light modulator *Opt. Eng.* **37** 942–947

[199] Cassinelli A, Chavel P and Desmulliez M P Y 2001 Dedicated optoelectronic stochastic parallel processor for real time image processing *Appl. Opt.* **40** 6479–6491

[200] Wu J-M, Kunzia C B, Hoanca B, Chen C-H and Sawchuk A A 1999 Demonstration and architectural analysis of CMOS/MQW smart pixel cellular logic processor for SIMD parallel pipeline processing *Appl. Opt.* **39** 2270–2281

[201] Shoop B L and Das P 2002 Mismatch-tolerant distributed photonic analog-to-digital conversion using spatial oversampling and spectral noise shaping *Opt. Eng.* **41** 1674–1687

[202] Kane J S 1998 Smart pixel feedforward neural network *IEEE Trans. Neural Networks* **9** 159–164

[203] Agranat A J *et al* 1989 *Opt. Lett.* **14** 1017

[204] Agranat A J *et al* 1992 *Opt. Lett.* **17** 713

[205] Agranat A J 1999 *IEEE-LEOS Summer Topical on WDM Components*

[206] Yablonovitch E 1987 *Phys. Rev. Lett.* **58** 2059

[207] John S 1987 *Phys. Rev. Lett.* **58** 2486

[208] Yoshino K *et al* 1999 Tunable optical stop-band and reflection peak in synthetic opal infiltrated with liquid crystal and conducting polymer as photonic crystal *Japan. J. Appl. Phys.* **38** L961–L963

[209] Leonard S W *et al* 2000 Tunable, two-dimensional photonic crystals using liquid crystal infiltration *Phys. Rev.* **B61** R2389–R2392

[210] Wallace J 2000 *Laser Focus World* **36**

[211] Lindvold R L and Lausen H 1997 Projection display based on optically-addressed SLM using bacteriorhodopsin thin film *Proc. SPIE* **3013** 202–213

[212] Reddy K P J 1997 Analysis of bacteriorhodopsin and its applications in photonics *Proc. SPIE* **3211** 2–13

[213] Piuno P A E, Krull U J, Hudson R H E, Damh M J and Cohen H 1995 Fibre-optic DNA sensor for fluorometric nucleic acid determination *Anal. Chem.* **67** 2635–2643

[214] 1996 Fibre-optic evanescent wave bio-sensor for the detection of oligonucleotides *Anal. Chem.* **68** 2905–2912

[215] Holst G, Glud R N, Kuehl M and Klimant I 1997 A micro-optode array for fine scale measurements of oxygen distribution *Sens. Actuators* **B38/39** 122–129

[216] Illic B, Czaplewski D and Craighead H G 2000 Mechanical resonant immunospecific biological detector *Appl. Phys. Lett.* **77** 450–452

[217] Rowe-Taitt C A, Hazzard J W, Hoffman K E, Cras J J, Golden J P and Ligler F S 2000 Simultaneous detection of six bio-hazardous agents using a planar waveguide array biosensor *Biosens. Bioelectron.* **15** 579–589

[218] O'Brien M J II, Perez-Luna V H, Brueck S R J and Lopez G P 2001 A surface plasmon resonance array biosensor based on spectroscopic imaging *Biosens. Bioelectron.* **16** 97–108

[219] Cunningham B, Lin B, Qiu J, Li P, Pepper J and Hugh B 2002 A plastic colorimetric resonant optical biosensor for multiparallel detection of label-free biochemical interactions *Sens. Actuators* **B85** 219–226

[220] Lundgren J S, Watkins A N, Racz D and Ligler F S 2000 A liquid crystal pixel array for signal discrimination in array biosensors *Biosens. Bioelectron.* **15** 417–421

[221] Efron U (ed) 1995 *Spatial Light Modulator Technology* (New York: Dekker)

[222] Collings N, Latham S G, Chittick R C and Crossland W A 1990 Reconfigurable optical interconnect using an optically addressed light valve *Int. J. Opt. Comput.* **1** 31–40

[223] Patel J S and Silberberg Y 1995 Liquid crystal and grating-based multiple wavelength cross-connect switch *IEEE Photon. Technol. Lett.* **7** 514–516

[224] Manolis I G, Wilkinson T D, Redmond M M and Crossland W A 2002 Reconfigurable multi-level phase holograms for optical switches *IEEE Photon. Technol. Lett.* **14** 801–803

[225] Wilkinson T D and Crossland W A 2001 Optical routing with LC arrays *Proc. SPIE* **4534** 64–69

[226] Crossland W A *et al* 2000 Holographic optical switching: the ROSES demonstrator *J. Lightwave Technol.* **18** 1845–1854

[227] Crossland W A, Holmes M J, Robertson B and Wilkinson T D 2000 Liquid crystal polarization independent beam steering switches for operation at 1.5 microns *LEOS 2000* vol 1 (Piscataway: IEEE) pp 46–47

[228] Apter B, Acco S and Efron U 2001 A study of LC-based sub-wavelength diffractive optical elements for optical cross-connect applications *Proc. SPIE* **4457** 20–30

[229] Kannie T *et al* 2002 A highly dense MEMS optical switch array integrated with planar lightwave circuit *Int. Conf. Micro-Electro-Mechanical Systems (MEMS)* (Piscataway: IEEE) pp 560–563

[230] Bakke T, Tigges C P and Sullivan C T 2002 1 × 2 MOEMS switch based on silicon on insulator and polymeric waveguides *Electron. Lett.* **38** 177–178

[231] Haven V S *et al* 2001 Recent advances in bubble-actuated cross-connect switches *CLEO/Pacific Rim—Tech. Digest* vol 1 (Piscataway: IEEE) pp I-414–I-415

[232] Chen D *et al* An optical cross-connect based on micro-bubbles *Micro-Electro-Mechanical Systems (MEMS), ASME Int. Mech. Eng. Congress, ASME 2000*, pp 35–37

[233] Walker A C *et al* 1998 *Appl. Opt.* **37** 2822–2830

[234] McCarthy A, Tooley F A, Laprise E, Plant D V, Kirk A G, Oren M and Lu Y 2000 Free space optical interconnect system using polarization-rotating modulator arrays *Proc. SPIE* **4089** 272–277

[235] Liu Y 2002 Heterogeneous integration of OE arrays with Si electronics and microoptics *IEEE Trans. Adv. Packaging* **25** 43–49

[236] Kasahara K M, Kim T J, Neilson D T, Ogura I, Redmond I and Schefeld E 1998 Wavelength division multiplexing free-space optical interconnect networks for massively parallel processing systems *Appl. Opt.* **37** 3746–3755

[237] Tuantranont A, Bright V M, Zhang J, Zhang W, Neff J A and Lee Y C 2001 Optical beam steering using MEMS-controllable microlens array *Sens. Actuators* **A91** 363–372

[238] Huang D *et al* 2001 Free space optical interconnection of 3-D opto-electronic VLSI chip stacks *Proc. SPIE* **4292** 95–104

[239] White H J *et al* 1999 Optically connected parallel machine: design, performance and application *IEE Proc. Opto-Electron.* **146** 125–136

[240] Weiner A M, Heritage J P and Kirschner E M 1988 *J. Opt. Soc. Am.* **B5** 1563

[241] Weiner A M 2000 Femto-second pulse shaping using spatial light modulators *Rev. Sci. Instrum.* **71** 1929–1960

[242] Wefers M M, Nelson K A and Weiner A M 1996 *Opt. Lett.* **21** 746

[243] Nuss M C and Morrison R L 1995 *Opt. Lett.* **20** 740

[244] Karasawa N, Li L, Suguro A, Shigekawa H, Morita R and Yamashita M 2001 Optical pulse compression to 5 fs by use of only a spatial light modulator for phase conjugation *J. Opt. Soc. Am.* **B18** 1742–1746

[245] Bleha W P 2000 D-ILA technology for electronic cinema *Dig. Soc. Inf. Display (SID) 2000* pp 310–313

[246] Van Kessel P F and Hornbeck L J 1998 A MEMS-based projection system *Proc. IEEE* **86** 1686–1704

[247] Sayyah K, Efron U and Forber R A 1995 Color-sequential crystalline Si-LCLV-based projector for consumer HDTV *Proc. Soc. For Information Displays (SID) Digest* p 520

[248] Morrissy J H, Pfeiffer M, Schott D and Vithana H 1999 Reflective microdisplays for projection or virtual-view applications *Soc. Inf. Display (SID) Digest* pp 808–811

[249] Gleckman P and Schuck M 2001 Optical characteristics of a high performance LCoS virtual display *Proc. Soc. Inf. Display (SID)* vol XXXII pp 62–65

[250] Spitzer M B, Zavracky P M, Crawford J, Aquilino P and Hunter G 2001 Eyewear platforms for miniature displays *Proc. Soc. Inf. Display (SID)* vol XXXII pp 258–261

[251] Peli E, Reed Hedges T, Tang J and Landmann D 2001 A binocular stereoscopic display system with coupled convergence and accommodation demands *Proc. Soc. Inf. Display (SID)* vol XXXII pp 1296–1299

[252] Haseltine E C 2000 Displays for location based entertainment *Proc. Soc. Inf. Display (SID)* vol XXXI pp 962–965

[253] Garibyan O V *et al* 1981 Optical phase conjugation by microwatt power of reference wave via liquid crystal light valve *Opt. Commun.* **38** 67–70

[254] Marom E and Efron U 1987 Phase conjugation of low power optical beams using liquid crystal light valves *Opt. Lett.* **12** 504–506

[255] Johnson K M *et al* 1990 High speed, low-power optical phase conjugation using a hybrid amorphous silicon/ferroelectric liquid crystal device *Opt. Lett.* **15** 1114–1116

[256] Shirai T, Barnes H and Haskell T G 2001 Real time restoration of blurred image with liquid crystal adaptive optics system, based on all-optical feedback interferometry *Opt. Commun.* **188** 275–282

[257] Neil M A A *et al* 2002 Active aberration correction for the writing of three-dimensional optical memory devices *Appl. Opt.* **41** 1374–1379

[258] Miller D T 2000 Retinal imaging and vision at the frontiers of adaptive optics *Phys. Today* January 31–36

[259] 1989 Recent advances in binary optics *Proc. SPIE* **1052** 85–90

[260] Sheppard C J R 1999 Binary optics and confocal imaging *Opt. Lett.* **24** 305–306

[261] Matic R 1994 Blazed phase liquid crystal beam steering *Proc. SPIE* **2120** 194–205

[262] McManamon P F *et al* 1996 Optical phased array technology *Proc. IEEE* **84** 268–297

[263] Suguira N and Morita S 1993 Variable focus liquid-filled optical lens *Appl. Opt.* **32** 4181–4186

[264] Commander L G, Da S E and Selviah D R 2000 Variable focal length microlenses *Opt. Commun.* **177** 157–170

[265] Paige E G S and Sucharov L O D 2001 Enhancement of the imaging performance of a variable-focus Fresnel zone plate based on a single binary phase-only SLM *Opt. Commun.* **193** 27–38

[266] Efron U, Wu S T, Grinberg J and Hess L D 1985 Liquid crystal-based, visible to infrared dynamic image converter *Opt. Eng.* **24** 111–118

[267] Wu S T, Efron U and Hsu T Y 1988 Near infrared to visible image conversion using a Si–liquid crystal light valve *Opt. Lett.* **13** 13–15

[268] Bertsch A *et al* 1997 *Photobiology* **A107** 275

[269] Chatwin C *et al* 1998 UV microstereo-lithography *Appl. Opt.* **37** 7514–7522

[270] Yariv A and Yeh P 1984 *Optical Waves in Crystals* (New York: Wiley) chapter 7

[271] Saleh B E and Teich M C 1991 *Fundamentals of Photonics* (New York: Wiley) pp 928–929

[272] Lipson S G, Lipson H and Tannhauser D S *Optical Physics* 3rd edn. (Cambridge: Cambridge University Press) pp 397–399

[273] Using the system impulse response function $h(t)$ theory, a commonly used stability criterion is that $\int |h(t)|\,dt$ must be finite, Oppenheim A V and Willsky A S 1997 *Signals and System* (New York: Prentice-Hall) p 114

[274] Butcher P N and Cotter D 1990 *The Elements of Nonlinear Optics* (Cambridge: Cambridge University Press) chapter 2

[275] Moss T S, Burrell G J and Ellis B 1973 *Semiconductor Opto-Electronics* (New York: Wiley)

[276] Efron U 1991 Liquid crystals; materials, devices and applications *Handbook of Microwave and Optical Components* vol 4, ed K Chang (New York: Wiley) pp 419–421

[277] Pepper D M, Gaeta C J and Mitchell P V 1994 Real-time holography, innovative adaptive optics and compensated optical processors using spatial light modulators *Spatial Light Modulator Technology*, ed U Efron (New York: Dekker) pp 627–631

Appendix
On the dynamic range–speed trade-off

The following derivation represents an intuitive argument rather than a rigorous proof and should be treated as such. We chose to include it despite its approximate nature, as the consequences of this hypothesis may be far reaching in terms of the fundamental limits of SLMs.

First let us generalize the term dynamic range of the system to be the range of values which the appropriate optical response function of the particular modulator system can take. It is the range of the electro-optic coefficient for an electro-optic modulator system or the range of the electro-absorption coefficient for the MQW modulator system. The speed of response is obviously directly related to the frequency bandwidth of the system. With this in mind, taking the electro-optic effect as an example, the change in the refractive index of the material Δn, upon the application of an electric field E, is given in general by [270]

$$\Delta n \sim n^3 rE \tag{A1}$$

where r is the appropriate electro-optic coefficient. This behaviour can be presented more generally, as a response of a system parameter $P(t)$, to an external driving field $F(t)$.

It can be shown that the temporal response function $\chi(t)$ (or the generalized susceptibility) of the system acts essentially as an impulse response function to the external driving field $F(t)$, such that:

$$P(t) = \int \chi(t-\tau)F(\tau)\mathrm{d}t \tag{A2}$$

where in the frequency domain the convolution operation of equation (A2) turns into a simple product, namely [271, 272]

$$P(\omega) = \chi(\omega)F(\omega). \tag{A3}$$

This is a generalization of equation (A1) above with P, χ, F, replacing Δn, r, E, respectively.

Now, the DC value of χ, namely, $\chi_{\mathrm{DC}} = \int \chi(t)\mathrm{d}t = \chi(\omega = 0)$, must be *finite*, based on the material stability considerations. This is so since an infinite value would imply that a vanishingly small fluctuation in the external field δF would result in a finite change of the system parameter P (e.g. the refractive index) [273]. However, since the latter is determined by the structure of the particular material system (e.g. an electro-optic crystal), a finite constant (DC) change in its value implies a finite structural change. Such a finite structural change, with a vanishingly small perturbation, indicates an inherent instability. Now, by the Parseval identity:

$$\left|\int \chi(t)\right|^2 \mathrm{d}t = \int |\chi(\omega)|^2 \mathrm{d}\omega \tag{A4}$$

and hence the finiteness of $\chi_{\mathrm{DC}} = \int \chi(t)\mathrm{d}t$ implies a finite value for $\int |\chi(t)|^2\mathrm{d}t$, and hence for $\int |\chi(\omega)|^2\mathrm{d}\omega$. Assuming a finite bandwidth, $\Delta\omega$, for $\chi(\omega)$, and making the approximation:

$$I\chi = \int \chi(\omega)\mathrm{d}\omega \sim \chi_{\mathrm{AV}}\Delta\omega \tag{A5}$$

where χ_{AV} is an average value for χ (ω) within the finite bandwidth $\Delta\omega$, we conclude that the finite value of $I\chi$ indicated by equation (A4) implies that the susceptibility χ_{AV} must, in general, be traded off against the system bandwidth, $\Delta\omega$, or speed of response. Thus, we conclude that there must exist, indeed, a trade-off between the dynamic range of the generalized electro-optic response function and the speed of response of that system.

Next, we need to consider the frequency range of the system, where there are basically two physically different frequency regions. The first, $\omega_1 = \omega_{\mathrm{RF}}$, is the RF field (e.g. the low-frequency externally applied electric field that controls the electro-optic effect). The second frequency region, $\omega_2 = \omega_{\mathrm{opt}}$, is the optical frequency range encompassing the spectral (wavelength) region in which this particular susceptibility (say, the EO effect) is nonzero. Typically, we have: $\omega_{\mathrm{opt}} \gg \omega_{\mathrm{RF}}$. Let us consider in detail the particular EO example with its associated second-order susceptibility, $\chi(2)$, related to the EO (Pockel) effect. Using results from nonlinear optics, we have for the second-order polarization $P^{(2)}(t)$ [274]

$$P^{(2)}(t) = \varepsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\omega_1 \int_{-\infty}^{\infty} \mathrm{d}\omega_2 \chi^{(2)}(-\omega; 0, \omega)E_1(0)E_2(\omega)\mathrm{e}^{\mathrm{j}\omega t} \tag{A6}$$

where we take: $\omega_2 = \omega_{\mathrm{opt}} = \omega$; $\omega_1 = \omega_{\mathrm{RF}} = 0$ (low frequency); $E_1(0) = \text{const}$ (=amplitude of the external bias field); $E_2(\omega) = E_2^{(0)} = \text{const}$ (=amplitude of the optical field). We now assume, as in the previous argument, that: $\int \chi^{(2)}(\omega)\mathrm{d}\omega \sim \chi^{(2)}_{\mathrm{AV}}\Delta\omega$ is finite over the finite (optical) bandwidth $\Delta\omega$ and thus we have an upper limit to the polarization given by:

$$|P^{(2)}(t)| \leq \Delta\omega_2\Delta\omega_1\chi^{(2)}_{\mathrm{AV}}E_1(0)E^0_2. \tag{A7}$$

now the polarization, or its variation with the applied field, is directly related to the variation in the electronic or ionic displacement $\langle x\rangle$ via [275]

$$P = Ne\langle x\rangle \tag{A8}$$

and so again due to material stability considerations P, or $P^{(2)}(t)$, must remain finite, given a finite $\langle x\rangle$, for any finite field levels below the breakdown limit of the material.

Therefore, the conclusion is that the quantity:

$$\mathrm{SBP} = \Delta\omega_2\Delta\omega_1\chi^{(2)}_{\mathrm{AV}} = \Delta\omega_{\mathrm{opt}}\Delta\omega_{\mathrm{RF}}\chi^{(2)}_{\mathrm{AV}} \tag{A9}$$

must be finite for this case of an EO material system.

So, we see that the GBWP constant takes on a new form for the 'susceptibility–bandwidth product' (SBP), which involves *two* frequency ranges rather than one, due to the two operating fields: the RF and the optical field.

We can conveniently express this new SBP, in units of frequency bandwidth (Hz) × optical bandwidth (cm^{-1}) × optical susceptibility, which we take as proportional to the linear electro-optic coefficient or $\Delta n/\Delta E$ ($\mathrm{cm\,V^{-1}}$). So, for the EO case we have the SBP quantity in units of $\mathrm{Hz\,V^{-1}}$.

It can be argued that the above considerations can be extended to a generalized susceptibility (e.g. AO, magneto-optic). This is so since the polarization modulation resulting from the application of the appropriate external field and with it, the associated ionic or electronic displacement, must remain below the structural damage level, regardless of the optical modulation mechanism (e.g acoustic or magnetic). Now, since this range of acceptable ionic or electronic displacement in solids, although not constant, varies within one to two orders of magnitude, we expect the SBP for all solid state material systems to be also roughly within one to two orders of magnitude.

Let us examine this hypothesis for three material systems:

Liquid crystals: $\Delta\omega_{\mathrm{RF}} \approx 10^4$ Hz (FLC); $\Delta\omega_{\mathrm{opt}} \approx 2.5\times10^4\,\mathrm{cm}^{-1}$ (400–10 000 nm); $\Delta n/\Delta E \approx 0.2/10^4$ ($\mathrm{cm\,V^{-1}}$). Here, we get for the $\mathrm{SBP_{LC}}$:

$$\mathrm{SBP_{LC}} = \Delta\omega_{\mathrm{opt}}\Delta\omega_{\mathrm{RF}}\chi^{(2)}_{\mathrm{AV}} \approx \Delta\omega_{\mathrm{opt}}\Delta\omega_{\mathrm{RF}}[\Delta n/\Delta E] \approx 5\times10^3\ \ (\mathrm{Hz\,V^{-1}}).$$

MQW modulator (electro-refraction): $\Delta\omega_{\mathrm{RF}} \approx 10^{10}$ Hz; $\Delta\omega_{\mathrm{opt}} \approx 30\,\mathrm{cm}^{-1}$ (850–852 nm); $\Delta n/\Delta E \approx 0.01/10^5$ ($\mathrm{cm\,V^{-1}}$). Thus we get for the $\mathrm{SBP_{MQW}}$:

$$\mathrm{SBP_{MQW}} \approx \Delta\omega_{\mathrm{opt}}\Delta\omega_{\mathrm{RF}}[\Delta n/\Delta E] \approx 3\times10^4\ \ (\mathrm{Hz\,V^{-1}}).$$

Solid-state electro-optic modulator ($r = 30\,\mathrm{pm\,V^{-1}} = 3\times10^{-9}\,\mathrm{cm\,V^{-1}}$): $\Delta\omega_{\mathrm{RF}} \approx 10^7$ Hz (FLC); $\Delta\omega_{\mathrm{opt}} \approx 1.5\times10^4\,\mathrm{cm}^{-1}$(400–1000 nm); $\Delta n/\Delta E \approx n^3 r \approx 30\times3\times10^{-9}\mathrm{cm\,V^{-1}}$;

$$\mathrm{SBP_{EO}} \approx \Delta\omega_{\mathrm{opt}}\Delta\omega_{\mathrm{RF}}[\Delta n/\Delta E] \approx 1.4\times10^4\ \ (\mathrm{Hz\,V^{-1}}).$$

We thus see that the SBPs of all these three markedly different material systems, with orders of magnitude variations in their RF frequency responses, spectral bandwidths and optical susceptibilities, come to roughly the same level, within an order of magnitude. This finding supports the 'universality' hypothesis for this quantity.

We also note that for systems with comparable spectral (optical) bandwidths such as liquid crystals and solid-state EO materials, the SBP constancy simplifies approximately to a trade-off between the EO susceptibility or $\Delta n/\Delta E$ and the RF bandwidth $\Delta\omega_{\mathrm{RF},}$ in analogy to the well-known GBWP.

Finally, an interesting situation arises when we can invoke a trade-off between the optical susceptibility and the frequency response within the *same* material system, e.g. by incorporating an optical feedback in a PA nematic LCLV [276, 277]. In this case, it can be shown that the product of the open-loop optical gain defined for the phase modulating LC-SLM by:

$$\Delta\Phi_{\mathrm{out}} = G_0 \Delta\Phi_{\mathrm{in}} \tag{A10}$$

and the LC frequency bandwidth,

$$\Delta\omega = 1/\tau_0 \tag{A11}$$

the GBWP G_0/τ_0, remains unchanged as the optical feedback mechanism is turned on, namely when we have:

$$\Delta\Phi_{\mathrm{out}} = G_{\mathrm{CL}} \Delta\Phi_{\mathrm{in}}$$

where for the closed loop system with a negative feedback fraction, β, we have:

$$G_{\mathrm{CL}} \approx \frac{G_0}{1+\beta G_0} \tag{A13}$$

and the associated LC frequency response is:

$$\Delta\omega = 1/\tau_{\mathrm{CL}}. \tag{A14}$$

It can be shown that [276]:

$$G_{\mathrm{CL}}/\tau_{\mathrm{CL}} \approx G_0/\tau_0.$$

This GBWP constancy allows a 50 Hz (open-loop) PA nematic LCLV to attain kilohertz frequency response using a negative feedback, closed loop system [277]. This is in close analogy to the constant GBWP behaviour in electronic systems (e.g. operational amplifiers).

C2.5
Organic electroluminescent displays

Nicholas Baynes and Euan Smith

C2.5.1 Introduction

The technology of organic light emitting devices (OLEDs) is a relatively new and radical development in the established field of displays. However, since the demonstrations of efficient light emission from small organic molecules [1] and conjugated polymers [2], their potential application in displays has been most compelling. These demonstrations built on earlier studies of organic electroluminescence, notably in polymers [3] and small molecules [4].

There are many attractive features of organic devices which draw them towards displays. Compared to inorganic semiconductors, organic devices appear to offer a clear route to large-scale manufacturing on cheap substrates with easy integration of different colour devices. Compared to liquid crystal displays (LCDs), OLEDs avoid the need for separate back lighting and for colour filters, enabling thinner and lighter construction with higher power efficiency. OLEDs also offer a good angular distribution of light, which is possible but complex for LCDs. Compared to plasma display panels, OLEDs offer smaller pixels so higher resolution. However, an OLED is not a panacea. For example, it will never offer lower power consumption than the reflective LCD used in a digital watch and may not attain comparable lifetime to the discrete inorganic LEDs used in very high-brightness stadium-sized displays. Nevertheless, there is an enormous and diverse middle ground in the display market that organic devices have the potential to capture on the basis of improved cost–performance trade-offs. Many significant applications are as yet unseen and will only emerge if and when the technology enables them. As recently as 1980, who would have predicted the present-day importance of LCDs to lap-top computers and mobile phones?

This paper aims to show how the science of OLEDs can be turned into the technology of useful displays. There are many similarities between small-molecule- and polymer-based technologies, so the two are covered together, with their differences only brought out as necessary. The basic device physics, the chemical and material properties of specific organic materials are not covered in detail here. For this, the reader is referred to chapter B11 and to the further reading list at the end of this paper.

The first half of this paper concentrates on display fabrication, explaining the methods, either available, or under development, for actually making the displays. Later sections look at displays from a system point of view, explaining the different types of display architecture and how they are driven.

Organic electroluminescence is a young technology. Small-molecule-based displays have been available in commercial products since 1999 (a Pioneer car radio was the first product and a Motorola mobile phone was released in 2001) and polymer-based displays are just emerging onto the market (an MP3 player and a Philips shaver became available late in 2002). Much of the detailed manufacturing know-how is proprietary, as well as still being subject to rapid improvements. However, enough is already published to allow a good introductory-level picture to be presented here. Because this is a rapidly developing field, we are reluctant to quote specific numbers on device performance (efficiency, lifetime, colour gamut, etc) as they will likely be rapidly out of date. For these the reader is referred to the websites of the main companies active in the field, which are given in the further reading section.

Organic light emitting materials divide into three structural categories: small molecules, polymers and dendrimers; and two light generation mechanisms: fluorescent and phosphorescent. These divisions are important for several reasons. The material structure determines how it can be applied to the substrate, solution processing for polymers and dendrimers, vacuum deposition for small molecules. This has significant implications for manufacturability discussed in section C2.5.6. The light generation mechanism determines the ultimate quantum efficiency, fluorescence being a quarter to a half the efficiency of phosphorescence [5]. These divisions are also reflected in the intellectual property portfolios of the companies leading the development of OLEDs with Cambridge Display Technology Ltd (CDT) owning fundamental patents related to conjugated polymer- and dendrimer-based displays and with Eastman Kodak Company and Universal Display Corporation (UDC) owning significant patents related to fluorescent and phosphorescent small-molecule-based displays respectively.

The display industry tends to discuss dimensions in inches. This is seen in both products (e.g. 15 in monitors, 42 in TVs) and processes (e.g. 6 in substrates, 14 in line). So both these figures and the metric equivalent (usually approximated as 1 in ~ 25 mm) are given. Unlike the round wafers used in the semiconductor industry, display substrates are usually square so when a substrate size is quoted as a single linear dimension the implication is this is the length of one side. Display size is usually quoted as the length of the diagonal. Substrates in-process are usually referred to as plates (not wafers!).

C2.5.2 Process flow example

This section describes how a simple OLED is made, to give a flavour of the processes and technology involved. As an example, we have chosen a monochrome passive matrix display, such as might be suitable for a mobile phone. It is a 'bottom-emitting' display that means it is designed to be viewed through the transparent substrate and transparent anode, whilst the cathode and encapsulation are opaque. A schematic cross-section of such a display is shown in figure C2.5.1.

The organic materials have sufficiently low electrical conductivity that lateral conduction away from the metallic electrodes can generally be ignored. This means that, for a monochrome display, there is no need to pattern the organic layer. Like LCDs, pixels can be defined by merely patterning the electrodes and light is emitted only where the anode and cathode overlap. This is shown in figure C2.5.2.

A small-molecule display of this type would normally use thermal evaporation for deposition of both the organic layers as well as the top electrode. We therefore describe a polymer display as this uses the additional, but technologically straightforward, method of spin-coating to deposit the organic layers.

In an R&D environment, such displays could be made one at a time, but in manufacturing it would be more usual to process larger substrates with many identical displays tiled across them.

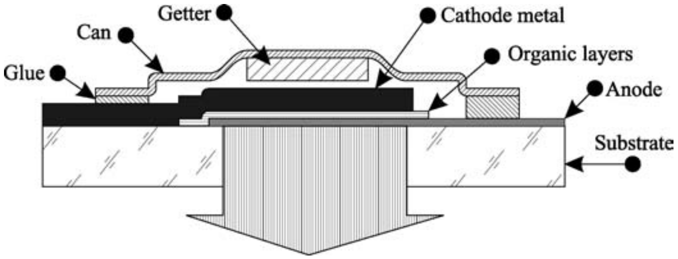


Figure C2.5.1. Schematic cross-section of an organic display. For clarity only one emissive pixel is shown.

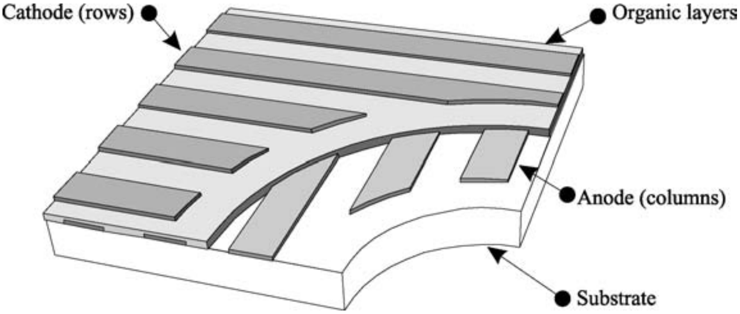


Figure C2.5.2. Schematic pixel layout in monochrome passive matrix display.

C2.5.2.1 Substrate preparation

Indium tin oxide (ITO) is generally used as the transparent conductor on the substrate and, in a simple display, its conductivity is high enough such that no additional metal layers are needed. The display manufacturer generally buys glass sheets, already uniformly coated with ITO. The desired pattern is imparted to the ITO layer using standard lithographic techniques involving a wet chemical etch.

To avoid short circuits between anode and cathode, it is vital that the organic layers are continuous. However, with organic layers typically less than 100 nm thick, it does not take a large piece of contamination to break the continuity of the layer. This means that to make a display with reasonable yield very thorough cleanliness must be maintained throughout the process. The cleaning of the substrate is therefore a critical process. This is typically done by a combination of mechanical agitation (brushing, ultrasonic and megasonic) in a series of liquids (water with detergent and then organic solvents). The final step in substrate preparation is an oxidizing surface treatment. This can be carried out either by oxygen plasma or UV–ozone treatment. It has the combined effects of removing any residual organic contamination (e.g. photoresist) and changing the surface energy of the ITO so that the first organic layer easily wets it.

C2.5.2.2 Organic layers

In a simple monochrome polymer display, two organic layers are spin coated. The first, known as the hole conduction layer, is usually an aqueous solution of polyethylenedioxythiophene (PEDT) and polystyrenesulphonic acid (PSS) with a large excess of PSS [6]. It is spin coated, to give a layer typically 70 nm thick, and then baked to speed up the drying process. Typical conditions are 10 minutes on a hot plate at 200°C.

The second layer is the emissive polymer, which is spun from an organic solvent, usually toluene or xylene. To achieve uniform appearance over the area of a display, the LEP layer must be of uniform thickness, typically 70 ± 5 nm. To achieve this level of uniformity on large substrates, a multi-stage spin coating process is used.

C2.5.2.3 Cathode deposition

The standard means of providing a metallic cathode contact is thermal evaporation in high vacuum, using a separate metal shadow mask in close contact with the display plate to define separate connections to each row of pixels in the display. LEPs are sensitive materials and it has been found that the simplest method of depositing metal without damaging the polymer is thermal evaporation with a low evaporation rate. If other techniques are used, such as electron beam evaporation or sputtering, great care must be taken to prevent exposure of the polymer to high-energy particles (electrons or ions), which would damage it. The injection of electrons into the LEP depends on intimate contact with a low-work-function metal, usually calcium or barium. However, these metals are invariably highly reactive. It is usual, therefore, to evaporate the minimum required thickness of the primary cathode metal and then provide a backing layer of a less reactive metal, usually aluminium or silver. The backing metal acts as a diffusion barrier and also provides a low resistance to lateral current flow. A typical cathode would therefore consist of a 10 nm thick calcium layer followed by a 200 nm thick layer of aluminium.

C2.5.2.4 Encapsulation and finishing

The active layers in a display require both mechanical and chemical protection. This is the job of encapsulation. The mechanical protection prevents damage during subsequent module assembly operations and is straightforward to provide. The chemical protection is to prevent attack by atmospheric gases, mainly oxygen and water vapour, and this is much tougher to provide.

The standard solution is to attach a metal or glass ‘can’ over the active display, using a UV-cure resin seal around the perimeter. The edge seal is the weak point and it will still allow some ingress. To reduce the impact on the display to an acceptable level, it is necessary to include a desiccant or ‘getter’ inside the package. This is a reactive material with a high surface area that captures most of the offending water molecules before they can attack the display.

After encapsulation, there are a variety of operations, the details of which are outside the scope of this paper. A typical sequence would be: test the displays, scribe and break the substrate into individual displays, attach electrical connections to the drive electronics and test again.

C2.5.3 Substrates

C2.5.3.1 Glass

Glass is the obvious choice of substrate, being used in all established display technologies from liquid crystal to CRT. In particular, its use in the liquid crystal industry means it is readily available coated with ITO. Its stiffness and dimensional stability make it straightforward to handle through the manufacturing process, and it can easily handle the required process temperatures. For highest conductivity, ITO requires processing above 400°C, which is too hot for many plastic substrates.

Glass is essentially SiO_2 with other materials added to reduce its melting point whilst ensuring that it remains amorphous (does not crystallize). There are very many different kinds of glass, but, for the relatively straightforward requirements of display substrates, the main contenders are standard sodalime and borosilicate glass. Sodalime is usually used because it is the cheapest but in displays it typically has a layer of SiO_2 deposited on its surface to prevent out-diffusion of sodium. Borosilicate has a higher melting point, does not have an out-diffusion problem and is tougher, particularly against thermal shock. However, for displays these advantages are not usually sufficient to justify the higher cost. The main disadvantages of glass in general are its brittle nature (displays have a tendency to break when dropped), its weight and its lack of flexibility.

C2.5.3.2 Plastic

One of the great promises of organic light emitters is to enable flexible, or at least thermo-plastically formable, displays. This has been a difficult challenge for liquid-crystal-based devices, because of the need to accurately maintain the thickness (of the order of 10 μm) of the liquid crystal film, which requires parallel-sided containment. The sensitivity to changes in this liquid-layer thickness can easily be seen by gently pressing on any LCD. Organic light emitters being entirely solid should not suffer from this problem.

The two main requirements on a flexible substrate are firstly to withstand the processing temperature, particularly the desired conditions for ITO deposition, and secondly to offer a sufficiently good diffusion barrier to oxygen and water. No plastic material can meet the barrier requirements on its own. This is well known in the food packaging industry where the shelf life of plastic-wrapped products (e.g. potato crisps/chips) can be significantly increased by using a metallized plastic bag. Organic displays are more sensitive to water than crisps (chips) and also require the barrier coating to be transparent. Suitable barrier-layer technology is being developed, but devices with lifetime close to those on glass have not yet been demonstrated. Single-layer barrier coatings of transparent materials, such as silicon dioxide, that are thin enough to be flexible, tend to suffer from pin holes, so the general principle is to apply a multi-layer stack of alternating barrier layers and plastic layers [7]. Care has to be taken to ensure low stress in the finished structure.

C2.5.3.3 Thin-film transistor

The practice of making LCDs with a simple electronic drive circuit at each pixel is well established, and used for practically all large LCDs, because it greatly improves many display performance parameters relative to a passive matrix design. The pixel drive circuits are usually made from thin-film transistors (TFTs) which are fabricated in a layer of silicon deposited on the glass. Silicon deposited in this way is not single crystal so the thin-film transistors cannot rival the performance of ones normally produced on a silicon wafer. However, for LCDs, which require no dc current, sufficiently good transistors can be made in amorphous silicon (which has charge carrier mobility up to $1\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$). For organic displays, such a low mobility is insufficient (a transistor capable of producing the current would be bigger than the pixel). Therefore, polycrystalline silicon—with mobility $40–120\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$—must be used. Amorphous silicon can be deposited at 300°C. There are two ways to make it polycrystalline. First, the whole substrate may be heated to at least 600°C. This is above the melting point of standard sodalime glass, so a more expensive material must be used as the substrate. In the second method a scanned laser rapidly heats the silicon with much less heating of the substrate. This is termed low-temperature polysilicon (LTPS). Polycrystalline silicon is becoming established in the LCD industry because it allows row and column drive electronics to be fabricated directly on the display substrate (rather than using separate ICs) and because it also allows much smaller transistors to be used at each pixel, giving a large aperture ratio. The large LCD manufacturers, such as Seiko Epson, Toshiba, Samsung, etc, who have developed polysilicon processing, are generally the only companies who have access to this kind of substrate, due to the cost and proprietary nature of the processes.

C2.5.3.4 Silicon

Silicon is a potentially attractive display substrate because of the ease of integrating the drive electronics. Silicon transistors can, of course, be made extremely small and so there is plenty of room to make much more complex electronics than the simple drive block found in TFT substrates.

There are two drawbacks to silicon substrates: cost and opacity. The high cost of growing large silicon wafers means that practical displays must be small, generally no more than 0.5 in (12 mm) across. This restricts them to close-to-eye applications such as camera viewfinders and virtual-reality goggles. The opacity of silicon means that the display must be top emitting. This is a drawback, as the technology for making a transparent cathode and transparent encapsulation is less well established than for the conventional opaque equivalents. However, if these technical problems can be solved top emission is a very attractive design.

C2.5.4 Electrode materials and metallization

C2.5.4.1 Anode materials

For displays emitting through the substrate, the most obvious requirements of the anode layer are sufficient transparency and conductivity. Inevitably, for a layer of a given material, there is a trade-off between these two properties controlled by the layer thickness, thicker layers being more conductive but less transparent. For a material with resistivity ρ (Ω cm) and optical absorption coefficient α (cm^{-1}) a layer of thickness t has a sheet resistance ρ/t and optical transmission $e^{-\alpha t}$. An appropriate figure of merit for the material, indicating its suitability as a transparent conductor, is therefore the product $\rho\alpha$, as this is the sheet resistance of a layer which would transmit 1/e of the light. There are other important criteria for anode materials. It must be possible to pattern the material, usually with a wet chemical etch. The anode material must also be sufficiently smooth to be covered by the organic layers. A localized material spike approaching the organic layer thickness can cause thinning of the organic layers. The resulting high local electric field causes increased current and accelerated degradation often leading to a black spot and potentially a short circuit. There is also concern that the anode material can act as a diffusion source of species that might degrade the device performance with time. The conductive metal oxides can be thought of as highly doped semiconductors—self-doped due to crystal defects partially caused by non-stoichiometry. The species in excess (e.g. indium in ITO) can diffuse out and interact with the organic layers. However, at the time of writing there is no consensus on this.

Although the work function of the anode material might be thought to be an important property, this is not critical when a doped conductive polymer such as PEDT is in contact with it. In this case, both anode and doped polymer exhibit metallic conduction with negligible barrier to charge flow regardless of work function.

ITO is almost universally used as the anode. It can achieve resistivity of $1.6 \times 10^{-4}\,\Omega$ cm and $\alpha \sim 10^4$ cm^{-1} although it does need annealing at 400°C to reach this. Standard ITO used in the LCD industry is not sufficiently smooth to make displays with a high yield, but flatter ITO is available from a number of suppliers. ITO can be etched in concentrated hydrochloric acid.

Other candidates for anode include zinc oxide and tin oxide. For a thorough review of transparent conductors see Hartnagel *et al* (1995) in the further reading list.

C2.5.4.2 Cathode materials

The primary requirement of the cathode metal is to inject electrons into the organic layers. Because the organic layers are generally not highly doped, this requires a metal with a work function similar to the LUMO (lowest unoccupied molecular orbital) of the organic layer. For polymers, this is usually calcium (work function 2.9 eV) or barium (2.7 eV).

Improved device performance has been reported for both small molecules and polymers, using an interface layer of lithium fluoride (or other group-one fluorides) between the organic layer and the cathode metal. These materials are insulators in the bulk, but act to enhance electron injection when less than a critical thickness (around 10 nm). The mechanism of this improvement is not fully understood but probably relates to interaction with both the organic layer and the cathode metal [8].

In a full colour display, the red, green and blue organic stacks will not necessarily present the same LUMO to the cathode. This is particularly likely for polymers because of the difficulty of making multilayer devices, whereas small-molecule designs can add an extra organic layer. In this situation, the optimum cathode may be different for the different colour sub-pixels. Whilst it may be technically possible to provide different cathodes for the different colours, the problems of cathode patterning explained below make this solution prohibitively expensive for practical manufacturing. Therefore, a common cathode is required that works sufficiently well with all three colours. At the time of writing, this is one of the unsolved problems for polymer devices. Adequate lifetime has been demonstrated separately in red, green and blue devices, but the blue devices used a different cathode.

C2.5.4.3 Additional metallization

In many display designs, additional metal tracking is needed to reduce the lateral resistance of the transparent anode. Without this, pixels a long way from the edge would experience a large voltage drop. This metal can be deposited and patterned using conventional semiconductor processing methods. It is usual to use aluminium with an adhesion layer of chrome.

C2.5.5 Patterning

As mentioned above, a simple monochrome display does not need any patterning of the organic layers. However, whenever a pixel matrix is used, patterning of the cathode is required and, for full colour displays, patterning of the organic materials and cathode is required. This is a significant issue because of the difficulty of using conventional photolithography once the organic materials and cathode metal are present. This difficulty is often explained in terms of the sensitivity of the organic materials, which is partly true, but in fact the biggest problem is due to the cathode metal which is very rapidly degraded by any contact with water. Conventional photolithography is not possible, because it uses aqueous developer.

C2.5.5.1 Cathode separators

For low-resolution displays, a separate metal shadow mask, held close to, or in contact with, the display, can be used to selectively deposit cathode metal. However, such shadow masks cannot easily be made with free standing strips of less than about 300 μm width and so this approach is impractical for pixels much smaller than 0.5 mm. The standard solution to this problem is to use 'cathode separator' structures on the substrate. These are made from a thick insulating layer patterned with an undercut so that the evaporated cathode metal is discontinuous over the edge as shown in figure C2.5.3.

C2.5.5.2 Laser ablation

For good encapsulation, the can needs to be well sealed to the substrate. However, displays made by spin coating the organic layers have polymer films covering the whole of the substrate. If the can were glued onto the polymer layers, the strength and impermeability of the seal would be severely compromised.

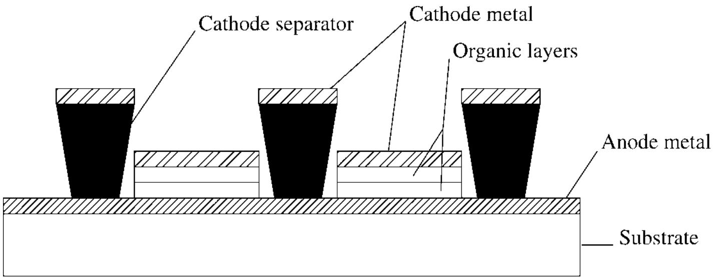


Figure C2.5.3. Cathode separator structures. In a passive matrix the anode lines run across the page and the cathode lines run into the paper. For small-molecule displays the separators also serve to pattern the evaporated organic layers. For polymers a separate patterning method is required for the organic layers.

Thus, there is a need for a low-resolution method of selectively removing the polymer layers. In research devices, this can be done manually with a solvent-soaked cotton bud. In some designs, it is possible to use the cathode metal as a mask and dry etch the polymers. However, the most suitable method for manufacturing is laser ablation. A high-power pulsed laser is fired at selected areas of the polymer. Laser energy is absorbed in the polymer which, by a combination of broken bonds and raised temperature, is then vaporized. With the right laser wavelength and power, it is straightforward to selectively remove the polymers without damaging the metal features below.

C2.5.5.3 Insulators

Almost all polymer display designs need some form of patterned insulating layer covering certain areas of the substrate. In spin-coated displays, this is required to cover up metal tracking and in ink-jet printed displays it defines the edge of the emitting pixels, making ‘wells’ into which the ink drops are deposited. In this function the insulator is also referred to as the ‘bank’ (as in river bank) and to work well it must be non-wetable by the inks. Fairly conventional materials (such as polyimide) and techniques are available from the semiconductor industry for making these insulator structures. The main requirement is a rounded edge profile so that the cathode metal covers it continuously.

C2.5.6 Organic deposition

C2.5.6.1 Vapour phase deposition

Thermal evaporation in a high vacuum is the standard method of depositing all small-molecule organic layers. It is a well established and reasonably straightforward process, which has the benefit of also being a purification process (most contamination will evaporate at a different temperature). Unfortunately, the only real way to pattern these organic layers is to do the deposition through a metal shadow mask held very close to the substrate. This is fine for simple low-resolution displays, but becomes increasingly difficult as the substrate size increases and the pixel size diminishes. The problem is caused by the need to maintain accurate alignment between the substrate and the mask under conditions of different thermal expansion coefficients and local source of heat. On a high-resolution active matrix display, the pixels may need to be deposited with positional accuracy of 10 μm across a substrate 50 cm wide. A steel mask plate, with a thermal expansion coefficient of $13 \times 10^{-6}{}^{\circ}\mathrm{C}^{-1}$, will move this much for a temperature change of only 2°C! Even if the mask is in thermal equilibrium with the glass substrate then, taking the thermal expansion coefficient of glass as $9 \times 10^{-6}{}^{\circ}\mathrm{C}^{-1}$, the differential expansion will cause similar misalignment after a temperature rise of about 6°C. The situation can be improved by matching thermal expansion coefficients, but nevertheless achieving a good yield on substrates which are large enough for economic manufacturing seems to be the largest problem facing small-molecule displays.

C2.5.6.2 Spin coating

Polymer solutions readily form stable solid films with good strength, uniformity and adhesion. This tendency is the basis of how paints, varnishes and simple glues work. The viscosity of the polymer solution stops it flowing too quickly off the substrate. As solvent leaves a polymer solution, the polymer chains form an interlocking network. Meanwhile, the forces of surface tension try to minimize the surface area, tending to give a flat or at least smooth film. It is this combination of effects that cause polymers to be often referred to as solution processable. This is in contrast to the behaviour of small molecules. Standard photolithography used in semiconductor processing exploits these polymer properties to make very uniform thin films of photoresist and it is this application that has led to the refinement of the spin coating process.

Because spin coating uniformly coats the whole substrate, the process is only applicable to simple monochrome polymer displays. In this application, compared to the alternatives (various printing techniques), the spinning process has the great advantage of simplicity. However, its big disadvantage is material wastage. For example, to coat a 6 in glass plate ($\sim 0.02\,\text{m}^2$), 10 ml of polymer solution is typically used. This solution contains 100–150 mg of polymer. If all of this polymer were to be spread out to a thickness of 70 nm, it could cover an area of about $2\,\text{m}^2$. Thus, in spin coating, only 1% of the polymer ends up on the substrate.

To achieve uniform appearance over the area of a display, the polymer films must be of uniform thickness. A typical specification would be 70 ± 5 nm. Whilst this is reasonably straightforward to achieve on a small substrate (e.g. 1 in, 25 mm), it is certainly not trivial for substrates of a manufacturing scale (>6 in, 150 mm). The standard solution is to use a multi-stage spin coating process, where more time is allowed for the wet film to reach equilibrium. This is usually achieved with a close-fitting cover over the spinning plate that retains a solvent-rich atmosphere. In some cases, the cover is spun at the same rate as the substrate.

All light emitting polymers are, at some level, susceptible to photo-oxidation, i.e. chemical reactions with oxygen or water that are greatly accelerated by light exposure. Because of this, processing is often carried out in a nitrogen or argon environment, typically inside a glove box. Depending on the polymer being used, this may not be necessary, so long as the exposure time to air is minimized and the lighting is filtered to cut out the higher-energy component that would be absorbed by the polymer.

LEPs are generally soluble in organic solvents, and these solvents have a very low ability to dissolve PEDT, so the LEP layer can be deposited on top of PEDT without seriously eroding the existing layer.

C2.5.6.3 Ink-jet printing

Ink-jet printing is the most promising method for manufacturing multicolour polymer displays [9]. As explained later, there are tighter constraints on properties of the polymer solution than for spin coating and typically two or more solvents are used to make a polymer ‘ink’. An ink-jet print head is essentially an ink-filled reservoir connected to a small aperture—the nozzle. In equilibrium, the ink does not flow out of the nozzle due to a combination of surface tension forces and the reduced pressure in the head resulting from having the main ink supply below the head. To actually print, a pulsed pressure wave is generated inside the reservoir and this causes a drop of ink to be ejected from the nozzle and propelled towards the substrate. There are broadly two methods of generating the pressure wave. In ‘bubble-jet’ a current pulse in a small electrical heater locally boils the ink. This only works well for aqueous inks so has not been successfully applied to polymer displays. The second method uses a piezoelectric actuator and this is much more suitable for solutions of light emitting polymers. For this process to work well in a given design of ink-jet head, the surface tension and particularly the viscosity of the ink must lie within narrow limits. As the total mass of polymer ejected in the drop will determine the final film thickness, the concentration of the ink is fixed by the drop volume and the pixel area. These combined constraints make developing good inks a complex process. There is a further complication relating to the requirement that, within each pixel, the polymer film must be flat over a good proportion of its total area. Generally, when drops of solution are allowed to dry on a surface, the resulting film is not flat. This is commonly seen when spilt coffee is allowed to dry leaving a dark perimeter, hence the result is often referred to as the ‘coffee ring effect’. The actual profile produced depends on a number of effects, namely the solvent evaporation rate across the drop surface, the speed of liquid flow within the drop and the speed of solute diffusion in the solvent. However, as a rule of thumb, low-boiling-point solvents tend to give a peak at the edge and high-boiling-point solvents give a peak in the middle. This means that a flat film can usually be achieved by using a blend of two appropriate solvents. Cross-sections of polymer pixels are shown in figure C2.5.4.

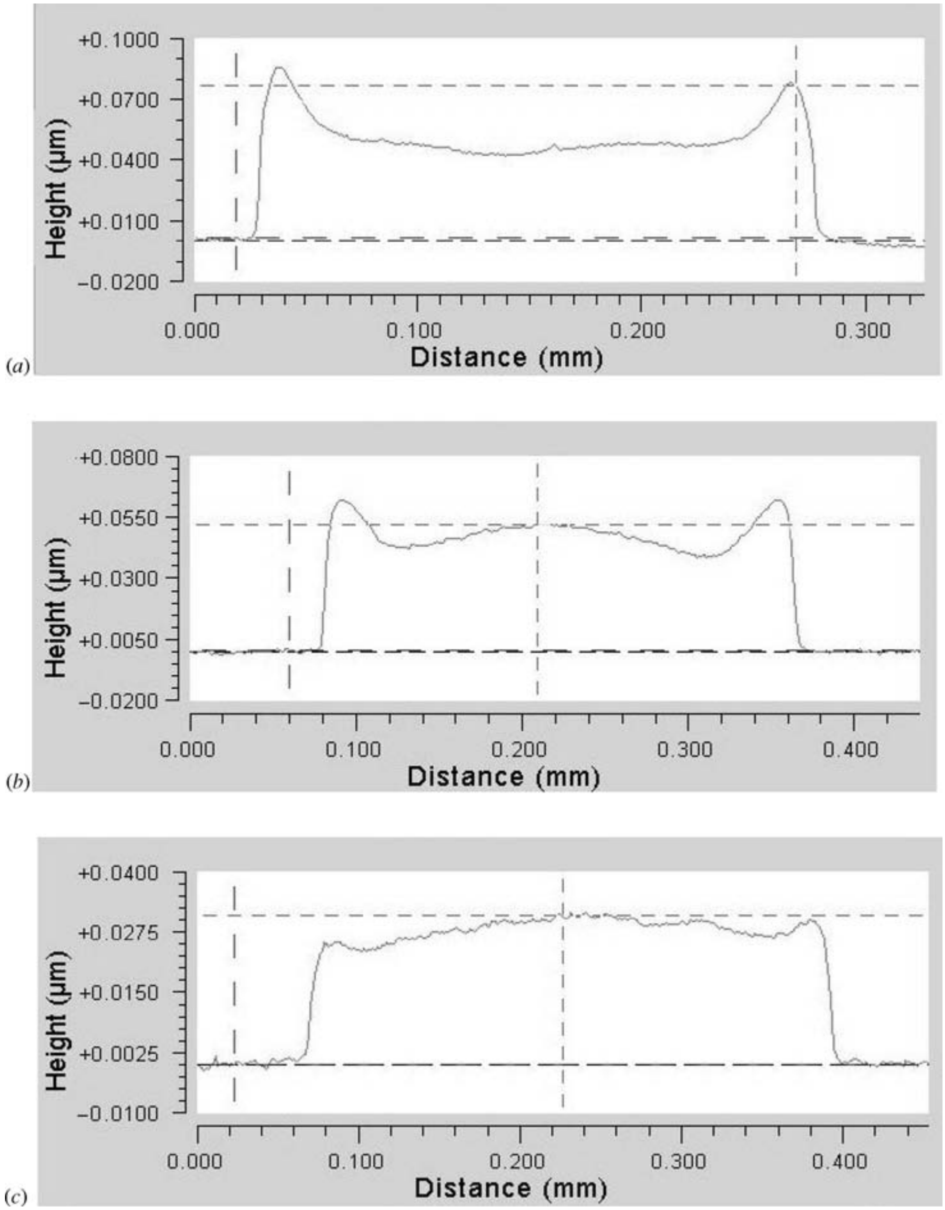


Figure C2.5.4. Cross-sections of ink-jet printed polymer films from a range of different solvent blends. Height measurements were taken using a white light interference scan. Solvents blends were A 90:10, B 70:30, C 50:50 (low boiling point:high boiling point). In this case the 50:50 blend gives the flattest film.

C2.5.6.4 Laser-induced thermal imaging (LITI)

LITI is an established method of patterning colour filters used in LCDs and has been demonstrated for patterning the organic layers in both types of display. A uniform organic film or stack is first prepared on a carrier substrate, one for each pixel colour. The organic stack is placed in contact with the display plate and selectively transferred using the energy from a pulsed laser. Patterning is achieved by masking the laser irradiation. The process is repeated for each colour. This process is promising from the point of view of manufacturing, but the problem remaining to be solved is how to minimize damage of the sensitive organic layers.

C2.5.6.5 Other printing techniques

Various other printing techniques, including screen printing, flexographic, gravure and micro-contact, have been proposed and demonstrated for patterning the organic layers in a display, although each has significant disadvantages. These generally relate either to the ink requirements that are difficult to meet for existing polymers (high viscosity and low evaporation rate) or to the risk of substrate damage from mechanical contact with the printer.

C2.5.7 Packaging

C2.5.7.1 Encapsulation

As mentioned earlier, both the organic light emitters and the cathode metal are seriously degraded by reaction with oxygen or water. It is therefore vital to provide encapsulation capable of keeping these gases out. Metals or glasses present an adequate diffusion barrier, so the standard method is to use the substrate itself as the front encapsulant and attach a metal or glass ‘can’ on the back of the display. The main issue is then sealing the joint between the can and the substrate. Various forms of glue can be used for this but none have sufficiently good barrier properties to make displays that can survive in a humid environment. The solution, which is a well-established technique for maintaining a good vacuum in isolated vessels such as cathode ray tubes, is to include a sacrificial desiccant or ‘getter’ material inside the package. The getter is a reactive material such as barium oxide, in a high-surface-area form (a powder). Any water molecule penetrating the edge seal will have a much greater probability of reacting with the getter than the device.

A potentially lower-cost solution to encapsulation is well established in the semiconductor industry. The ideal is to provide the required diffusion barrier by vacuum deposition of a thin conformal hermetic coating. Silicon nitride, aluminium oxide and aluminium nitride are examples of suitable material. Mechanical protection is provided separately by a moulded plastic package.

C2.5.7.2 Electrical connections

The finished display needs to be connected to drive circuitry. There exist various ways of doing this, all developed for other flat display technologies, but the details are beyond the scope of this paper. The main point to note is that any process involving high temperature can be problematic, as it can cause changes in the morphology of the organic layers, hence degrading the display performance.

C2.5.7.3 Black spots

A degradation phenomenon, where initially small circular areas within a diode fail to emit light, is commonly observed in OLEDs. These non-emissive areas are usually observed to grow with time. Growth of black spots is known to be caused by ingress of some reactive species, typically oxygen or water, through pin holes in the cathode [10]. In the case of encapsulated devices this reactive species must either originate inside the encapsulation, for example out-gas from an epoxy resin, or have permeated through the encapsulation. The exact nature of the interactions between the atmosphere and the device structure which lead to the decrease in light output are still controversial, and probably depend on the detail of the device construction [11].

The presence of pin holes in the cathode metal is due to contaminating particles that could arrive on the plate at any stage in the fabrication. Because the cathode metal is deposited in high vacuum, the metal atoms travel in straight lines from the small source and so any particle much larger than the organic layer thickness (~200 nm) will cause a break in the metal layer. Although it might be possible to close up these breaks by adding an extra metal layer deposited by a less directional technique such as sputtering, it is probably preferable to prevent the introduction of particles in the first place.

C2.5.8 Electro-optic response

OLEDs are thin, large-area LEDs. This simple statement points towards the most important properties of an OLED display element that govern how one would drive an OLED display, i.e. an LED requires current to emit light and is, most appropriately, current controlled. That it is a thin-film large-area structure means that the transport time through the device is fast. However, it also means that it has a significant capacitance and is prone, through any defects, to be a leaky diode. The total electro-optic response of the OLED device is split into three sections—electrical response, optical response and ageing effects.

C2.5.8.1 Electrical response

The structure of an OLED is a stack of various organic and inorganic layers. The three important layers, as far as the electrical properties are concerned, are the hole transport layer, emission layer and electron transport layer or, in the case of polymer systems, the cathode. The two layers either side of the emission layer have, typically, a much higher conductivity than the emission layer itself, and the device as a whole should only conduct when sufficient forward bias is applied, such that carrier injection into the emission layer can occur. Such an idealized device structure gives rise to the equivalent electrical circuit of an OLED of a diode and capacitor in parallel. The fabrication of real, slightly less than ideal, devices also gives rise to other important properties, in particular reverse-bias current leakage.

Capacitance

When below the threshold voltage in forward bias, and when reverse biased, the OLED structure reduces to a thin insulator ($<100\,\mathrm{nm}$) sandwiched between two large-area conductors, i.e. the perfect structure for a capacitor with a typical capacitance of the order of $300\,\mathrm{pF\,mm^{-2}}$. While under steady-state operation this property matters little; however, when pulse drive is required (see the later section on passive matrices) the capacitance becomes very important. It might be argued that, if this is a problem, the capacitance could be reduced by increasing the thickness of the emission layer. Increasing the emission layer thickness does indeed reduce the capacitance. However, it also increases the drive voltage of the diode. As the energy required to charge the capacitance varies as the voltage squared, a thicker device will tend to have an increased power consumption.

Diode properties

Similar to an inorganic LED, an OLED device possesses a well-defined threshold voltage, and above this the current flowing through the device turns on rapidly. As with an LED, the threshold voltage is related to photon energy of the emitted light; therefore, red devices possess a lower threshold voltage than blue. For polymer OLED devices turn-on voltages are typically in the range ~2 to ~3.5 V and in general the turn-on voltages for molecular materials are intrinsically higher than that of polymer (e.g. ~5 V when compared to ~2.5 V for a green polymer). It should also be noted that, in general, the threshold voltage can reduce at higher temperatures. However, unlike the LED, the (steady-state) current–voltage response of the OLED is not exponential—at least, not once there is a significant current flowing.

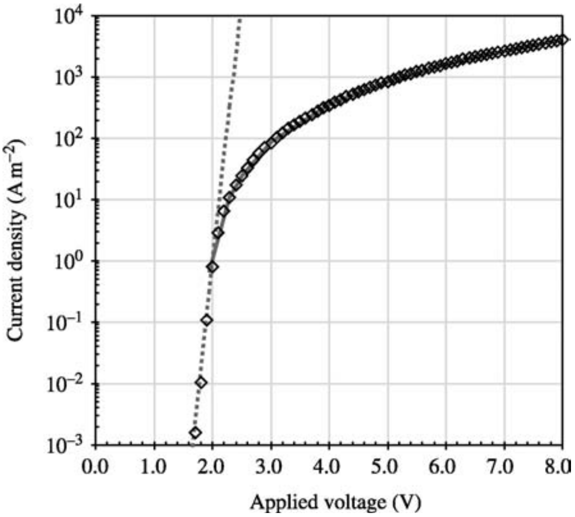


Figure C2.5.5. A typical *I–V* curve for an OLED device (diamonds) with an exponential fit of current density (dashed line) for low fields and a power law fit (solid line, see text) for high fields.

Figure C2.5.5 shows a typical *I–V* curve for an OLED device. Under low fields, the device is largely injection limited and this does indeed give rise to an exponential response. However, once the voltage is taken a little above threshold, the *I–V* response follows an approximate power law, similar to an old-fashioned valve diode, due to the current flow becoming space charge limited. Unlike a valve diode, the power law is not necessarily quadratic but can be anything from a power of 2 to 4 due to the carrier mobility being field dependent. The full *I–V* properties of an OLED device can be described by using the Murgatroyd equation [12]. However, over a specific region of interest the response can usually be simplified to $J \propto (V - V_t)^n$, where J is the current density and V the applied voltage, and V_t is the threshold voltage.

Reverse leakage

An OLED is an exceptionally thin large-area structure. Such a structure is inevitably susceptible to the effects of any defects, particulates or unusually thin areas of organic material, all of which can give rise to current leakage and, in the worst case, soft and hard shorts. A hard short will prevent a display element from functioning. A soft short will often burn out leaving a small non-functioning area of the pixel. Current leakage (which can also be the after-effect of a burnt-out soft short) is more benign but can cause problems for certain display types. This problem is termed ‘reverse leakage’ because it is only under reverse bias that leakage effects become noticeable, and a problem. Obviously, it is a process issue to reduce these effects to an acceptable level. However, some material sets also exhibit poor reverse leakage properties and so must be avoided (for passive matrix displays in particular, as these require good rectification ratios). Figure C2.5.6 shows the *I–V* characteristics of a number of similar devices, two of which show poor reverse leakage properties.

C2.5.8.2 Optical response

Light output—device driving

The current efficiency of a device is usually expressed in units of cd A^{-1}, and is typically slowly varying against drive level. This results in a light output that is, over a normal operating range,

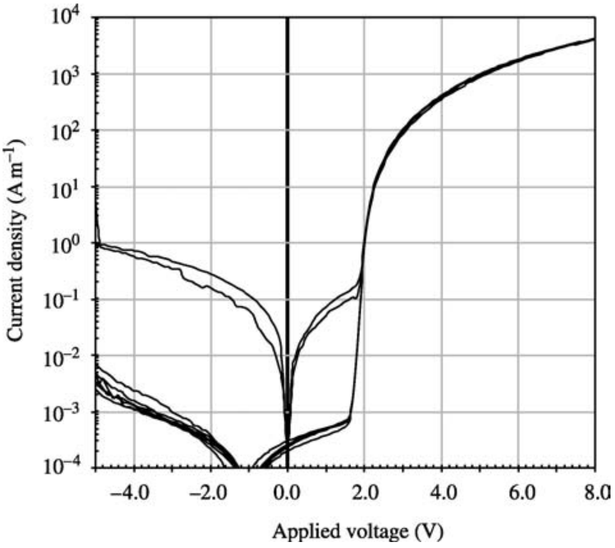


Figure C2.5.6. *I–V* characteristics of a number of similar devices, six with good and two with poor reverse leakage characteristics. The voltage offset, at low currents, of the good devices is a measurement artifact.

fairly linear with current. If this response is contrasted against the light output with voltage, it is clear that in order to obtain accurate control over a brightness level, current drive would be the preferred method of control. Further, even when only a single fixed brightness level is desired, the total luminous output of a voltage-driven pixel can vary considerably with device thickness, temperature, active area, series resistance, voltage threshold, etc, with the further complication that a voltage-driven pixel will age faster. Current drive offers much less sensitivity to such effects (than voltage drive) and thus offers a greater degree of uniformity over the display, between displays and over time. Figure C2.5.7 shows the light output against voltage and current for three nominally identical devices, demonstrating the reproducibility (and linearity) of current drive.

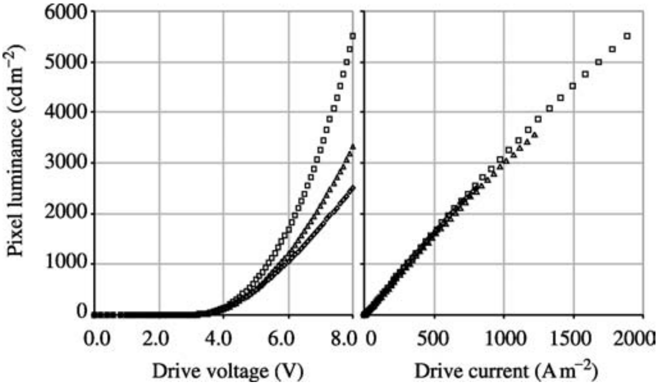


Figure C2.5.7. Light output against voltage and current for three nominally identical devices, demonstrating the reproducibility of current drive and its superiority over voltage drive.

Response time

This is really a 'non-issue' for OLED devices. The specific response time varies with materials, device geometry and applied field, but is typically of the order of 10 ns–1 μs, i.e. essentially instantaneous as far as a display application is concerned.

Light output coupling

For any light propagating within a high-refractive-index medium being incident on an interface with a low-refractive-index medium, there is a propagation angle to that interface beyond which the light is totally internally reflected (TIR). If light is generated with uniform emission in this high-index medium, a certain fraction of this light will be trapped within the medium due to TIR. Although the situation is complicated by dipole orientation etc, this picture describes the case for both small-molecule and polymer LED devices. In the SMOLED case, the fraction of light coupling out of a device is of the order of 17%, and for the polymer case it is ~30% [13]. Some potential methods to deal with this light loss are described in the section on system enhancements.

C2.5.8.3 Ageing effects

As a device is driven, there is a gradual loss of performance due to degradation of some part of the functional elements of the device. This manifests itself in three ways, an increase in the voltage required to achieve a given current density, a reduction in the light output for a given current density and, for some material systems, a shift in colour with time.

Usually the only quoted lifetime figure is a luminous intensity half-life for a given current drive or initial brightness; however, voltage and colour ageing can be as important for some applications. In particular, even if a device has a huge half-life, the voltage increase can seriously reduce the system lifetime if there is a requirement on a tight voltage rail (see figure C2.5.8).

In all cases it is the system life, determined by the driving conditions and the minimum acceptable display performance, which should be calculated to compare potential material sets. This is particularly the case for RGB display systems where the luminance half-life is often not the most relevant figure, as it is differential ageing, both between the colours and between different areas of the screen, which is often critical. Furthermore, in comparing different options of available materials, it is important to calculate the system lifetime of each complete RGB material set rather than taking one component in isolation.

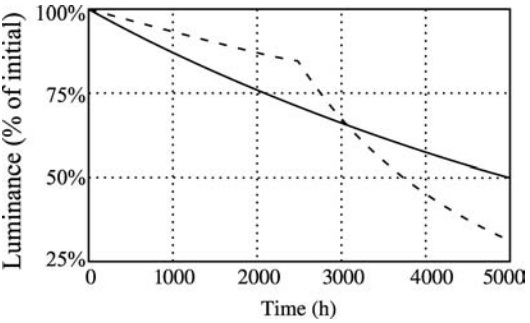


Figure C2.5.8. Simulated decay curves for two devices. Device 1 (solid line) has a luminance half-life of 5000 h when constant-current driven and a voltage change of 0.1 mV h^{-1}; for device 2 (dashed line) the figures are 10 000 h and 0.4 mV h^{-1}. A 1 V supply rail overhead is assumed.

A common example of this is the choice of blue material. It is typically the case that deep blue emitters will exhibit a shorter lifetime, at a given brightness, than lighter blue alternatives. However, the specific colour coordinate will determine the luminous intensity required to produce an acceptable white point. So, although a deeper blue may have a shorter lifetime, at a given luminous intensity, than a lighter blue, in an RGB display the deeper blue will not need to be driven as hard as the alternative and, therefore, will have a longer lifetime, when compared to the lighter blue, than the headline figure might suggest. Furthermore, as this is an effect of colour mixing to achieve the white, the colour coordinates of the red and green emitters will also modify the result, as well as the use of a deeper blue increasing the required luminosities of the other two emitters, shortening their respective lifetimes.

A discussion of the potential causes of ageing effects is beyond the scope of this review, and a quote of 'current' specifications, considering the pace of development, is liable to be seriously out of date by the time of reading, demonstrated in figure C2.5.9 showing the progress of blue polymer lifetime. However, the reader needs to be aware that ageing of a device depends strongly on drive level, duty cycle and environmental conditions (particularly temperature). Furthermore, it is not always possible to extrapolate reliably the lifetime under one set of conditions to another. It is therefore important to obtain the ageing characteristics important to an application for the conditions specific to that application. For current lifetime specifications, the reader should refer to announcements from the main developers of the different 'flavours' of OLED material (see company details at the end of the paper).

C2.5.9 Passive matrix displays

A passive matrix is the simplest possible matrix display and consists of arrays of rows and columns forming pixels where they intercept. One row is selected and all the pixels on that row are driven from the columns. The other rows are selected in turn until the entire frame has been scanned. From their inception OLEDs were, due to their diode nature, touted as ideal for the fabrication of passive matrix displays, with visions of huge screens for very low cost. Of course, life is never that simple. The limits to passive matrix display size, as shall be seen, are the requirement for current to emit light and, in particular, the effects of device capacitance.

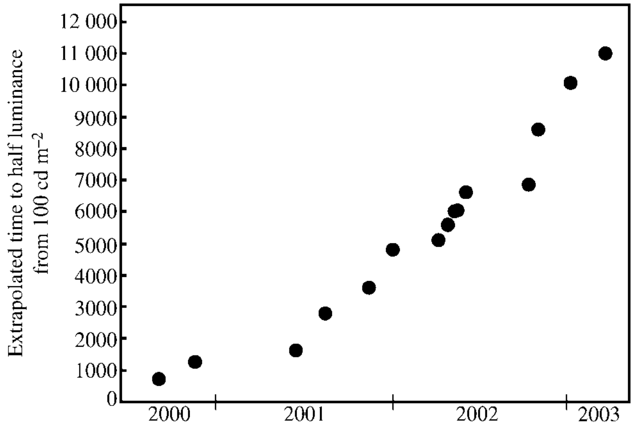
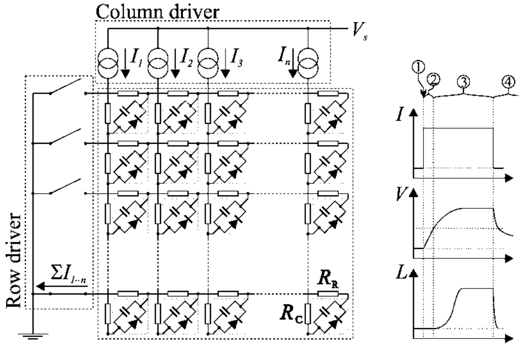


Figure C2.5.9. Announced blue polymer lifetime figures from CDT against announcement date.



(*a*) (*b*)

Figure C2.5.10. (*a*) Functional diagram of an OLED passive matrix. (*b*) Current, voltage and light output versus time of one pixel during a line scan.

C2.5.9.1 Passive matrix operation

Figure C2.5.10 shows (*a*) a functional diagram of an OLED passive matrix and (*b*) the current, voltage and light output of one pixel during a line scan. The line scan sequence (assuming current drive) is as follows, with the step numbers corresponding to the labels in figure C2.5.10(*b*);

(1) The row is selected (usually switched to ground) and the current, I, is driven into the column.

(2) The voltage on the device will increase linearly until the threshold voltage is reached. The time this will take is $t_C = CV_t/I$, where C is the capacitance of the column and V_t the threshold voltage.

(3) The pixel is now emitting light and the voltage gradually tends towards the nominal drive level. The time constant for this is harder to derive as it depends on the specific I–V relationship of the device.

(4) The current supply is turned off and the voltage on the device rapidly drops to the threshold voltage, and then very gradually drops back to 0 V, assuming no other pixels are driven.

Electrically, to complete the picture, the resistive losses from the pixel's row and column, plus the driver compliance, need to be added. Taking all of these into account, a simple model can be constructed which, despite its simplicity, can be a powerful predictor of the expected display performance.

Figure C2.5.11 shows results from an OLED passive matrix model, predicting power consumption per pixel versus number of rows. This is an iterative model based on the electrical response of an OLED pixel as described in section C2.5.8. These results show the potential dominance of capacitive power consumption in larger displays. With more rows comes more column capacitance and more charging cycles leading to an approximately quadratic increase in capacitive losses with row number. Furthermore, there is the knock-on effect of an increased current demand causing more resistive losses and driver compliance losses. It is the capacitive losses which often limit the practical size, resolution and/or brightness of a passive OLED display and it is clear that, if an efficient display is desired, control of capacitive losses is essential to the design of a passive matrix drive scheme.

C2.5.9.2 Leakage and defects

All the possible effects, on display operation, of current leakage (localized or over the display) and other display defects would depend strongly on the details of the specific implementation of the drive scheme.

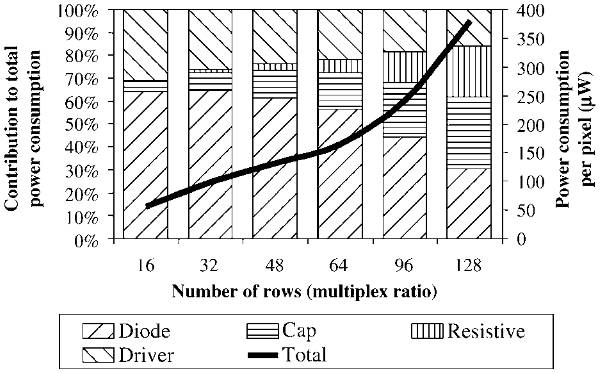


Figure C2.5.11. Power consumption per pixel versus number of rows in a passive matrix display. Results are calculated from an analytical model.

The simplest possible driver with no grey scale, an off-state anode clamped to 0 V and open circuit off rows, is relatively unaffected by display defects. Pixel shorts will, largely, only prevent the operation of that specific pixel and reverse leakage will, if not excessive, cause a slight loss of brightness of the pixel being driven (assuming current drive is used). With more significant leakage spatial cross-talk, apparent as streaking along the row and column from the active pixel, can become increasingly visible, although the acceptable threshold for this depends upon the application and displayed images. The opposite extreme might be a driver that uses methods to actively drive the deselected rows and columns off to minimize the effects of capacitance and improve contrast and grey scale but where a shorted pixel would cause all other pixels on a common row or column to cease functioning.

The selection of a driver for a display needs to be informed not only by the target application but also by the level and type of defects likely to be present in the display and the sensitivity of the specific drive scheme to such defects. A high-fidelity grey scale and contrast capable driver can only work well with a very low-defect-rate display—if the display in question tends to suffer from defects of any form then, in general, the simpler the driver the better.

C2.5.9.3 Grey-scale methods

There are two ways to modulate the quantity of light emitted by a pixel during a line scan — by changing the height of the current pulse (analogue drive) or its duration (PWM or pulse width modulation). The first of these might seem the obvious of the two, particularly as the light generation efficiency usually decreases with drive level; however, there are many more factors to bring into play, all of which are influenced by three factors—current density, capacitance and driver implementation.

The first of these is simply that, with analogue drive, the pixel is driven, on average, with a lower current density. Typically, this will result in more efficient operation (due to a lower drive voltage) leading to a longer lifetime. Knock-on effects might be the possibility of a larger, brighter or higher-resolution display becoming possible. This is tarnished slightly when it is realized that the benefits of a lower drive voltage are lost if the supply voltage must still hold to the highest potential voltage. However, at least this power dissipation is no longer on the panel itself, so many of the stated benefits still apply. Overall this factor is a vote for analogue drive.

The effects of capacitance paint a less clearly defined picture. PWM always has a full charge–discharge cycle per frame whereas, with analogue drive, the charging and discharging is always only the difference in drive level between pixels, resulting in another power advantage for analogue. However, this also results in the initial state of the voltage on a pixel being dependent (for analogue drive) on the state of the previous pixel causing potential cross-talk in the form of vertical edge blurring. The preferred choice here will very much depend on the target application.

Unfortunately, the real killer for analogue drive is the requirements on the column driver. In a typical application 64 grey levels might be required. These grey levels should, of course, have a gamma of ~2, therefore a dynamic range of 4096:1 is required. If this were all then the problem would be minor; however, as this is a light emissive display the drive scheme also needs to have a global brightness adjustment, typically of 100:1, which increases the dynamic range for analogue drive to ~400,000:1. In this case, PWM can, essentially, cheat, by separating the two adjustments—the grey levels are implemented by PWM but the global brightness, which does not need to be an accurate control, is controlled via the reference drive current. There are other factors in the design of the drive circuits. However, they all tend to point in favour of PWM also, so the vote here falls quite firmly on the side of PWM.

C2.5.9.4 Multiplex control

There are a number of controller enhancements which can be achieved through multiplex control, and all take advantage of the graph in figure C2.5.11—the lower the multiplex rate, the lower the power consumption.

Dual scan

The simplest and most effective method of making a display more efficient involves splitting the columns across the centre of the display and driving them from top and bottom, in essence splitting the display into two passive matrices. This comes at the cost of an increased component count (double the column drivers, although they need handle less than half the power) and possible lower yield (double the column connections to make). Nevertheless, this is a valuable method in achieving larger displays, the direct benefits of which can be seen, again, in figure C2.5.11, where the power consumption per pixel will be for half the multiplex of normal (e.g. a 100-line single-scan display would have almost double the power consumption of an equivalent dual-scanned display). With a dual-scan drive scheme, the largest display format with a passive matrix realistically possible in the near future increases to QVGA format (320 × 240 pixels).

Selective scan and screen-saver modes

Both of these schemes reduce the effective multiplex rate by driving a lower number of rows. In the case of a screen-saver mode this is a fixed reduction to a sub-set of the display (e.g. the central quarter in a phone display to show the number called), and for a selective scan this is an on-the-fly operation not to drive any rows not currently in use (e.g. the blank rows in between lines of text). Only driving (say) 32 rows out of 64 does not lower the power consumption figures down to those of a 32-row display, as the entire column capacitance is still present; however, significant savings (as well as an increased display life) can be made as is shown in table C2.5.1. It should be noted, however, that the full power savings will not be realized unless the power supply is adaptive and can adjust the supply voltage to minimize compliance—without this much of the power saved in the display panel will instead dissipate in the driver.

C2.5.9.5 Capacitance charging

Pixel capacitance can cause three effects: a power loss, an offset in the light output to drive level and cross-talk through residual capacitance. The power loss has been dealt with previously, the offset in light output is due to the time taken to charge the pixel up to the operating voltage and the cross-talk is due to the residual charge left at the end of a line scan (and present at the start of the next).

Table C2.5.1. Display power consumption with selective scan and screen saver modes.

Display mode	Standard PWM driver (mW)	Line skipping driver (mW)
All pixels on full brightness	188	188
Screen of text	41	36
Screen saver mode	12	5

Results from modelling a 96×64 monochrome green polymer display, $250\,\mathrm{cd\,m^{-2}}$ average luminance. Text screen assumes 25% of pixels on, 87.5% of lines active. Screen saver uses central 16 lines, 25% of those pixels on.

The simplest way to minimize the power loss is simply to do nothing, as corrective actions for the other two effects tend to increase power consumption. However, for some displays the cross-talk, in particular, can cause problems, and in this case some form of pre-charge is often used.

Pixel pre-charging is typically implemented via the application of a voltage to the pixel column, prior to the current drive signal, which charges the pixel to a set voltage. This injects current more rapidly than the current drive would be able to and thus keeps the charge-up ‘dead’ time to a minimum. This step is also insensitive to any residual charge on the pixel. The difficulty is deciding to what voltage the pixel should be charged.

The minimum voltage worth considering is the threshold voltage; the maximum is the nominal drive voltage; however, these two extremes have problems. Charging to threshold helps reduce cross-talk effects; however, it makes little difference to the offset in light output as there is still a lot of charging to go to reach the nominal drive level and the current driver is slow in charging this up. Pre-charging to the drive level can cause problems with the lowest grey levels as there will have been a quantity of light emitted prior to the current drive phase which will offset the light output, as well as the requirement to discharge at the end of the drive pulse to ensure excess light is not emitted at the end of the pulse.

The voltage selection would typically lie somewhere between these two extremes, and the sort of criteria used to select it might be that the light emitted during the charge-up period is no more than that required by the minimum grey level (for six bits of gamma 2 grey level, this would be 1/4096 of the maximum emission). This still has the requirement of a discharge cycle at the end of the line scan which is, of course, what increases the power consumption of the display when pre-charge is used.

Potentially it may be possible to use charge recycling methods, such as are used with LCD displays, to reduce the power consumption of an OLED passive matrix. The limiting factor in this case is the ITO resistance. If Q is the charge, R is the ITO resistance and t is the time that is available to recover the charge then resistive power loss during recovery is Q^2R/t. This means that, although such recycling could help reduce power consumption, in large high-resolution displays, where capacitive losses are greatest, there is also the highest ITO series resistance and least time in which to recover charge, so the benefit of charge recycling would be limited unless thicker ITO or anode bus bars were used to reduce the series resistance.

C2.5.10 Active matrix displays

It is clear that OLED passive matrices, while providing high-quality image reproduction for small display formats, are unsuitable for expansion to larger display formats, and that an active matrix scheme

Table C2.5.2. Mobilities of available transistor technologies.

TFT technology	Majority carrier mobility ($\mathrm{cm^2\,V^{-1}\,s^{-1}}$)
a-Si (amorphous silicon)	≤ 1
LTPS (low-temperature poly-silicon)	10–100
CGS (continuous grain silicon)	300–500
Bulk silicon	~1000

is required. The general case is, of course, the same for LCDs. However, as with passive matrices, there are particular factors that need to be considered when applying active matrix technologies to OLEDs, the primary ones being TFT (thin-film transistor) carrier mobility, TFT threshold voltage non-uniformities and OLED non-uniformities.

Table C2.5.2 shows typical carrier mobility for the various TFT technologies available, along with crystalline silicon for reference. Figure C2.5.12 shows the driver channel width requirements of a transistor to drive an OLED pixel at 300 cd m^{-2}. The $>$100 μm driver channel width requirement of a-Si does not rule it out as a potential TFT technology; however, it certainly makes the case borderline.

TFT threshold voltage variation is an issue with certain technologies and can, if not compensated for, cause significant brightness non-uniformities in a display. Of course variations in the OLED pixel response can also cause non-uniformities which, although not dependent upon the TFT technology directly, is dependent upon the design and operation of the pixel circuit which in turn is limited by the choice of technology.

There are a number of circuit types (and variations thereof) that can be used to drive an OLED pixel, and these are covered in the following sections. However, there is one other design factor important to the operation of an active matrix display and this is the choice of top or bottom emitting structure.

The standard OLED structure uses a transparent anode (e.g. ITO) and reflective metallic cathode (see figure C2.5.13*a*). This means that the device must emit light through the substrate and therefore, as with LCDs, the aperture ratio must be shared between the pixel circuitry and emitting area. However,

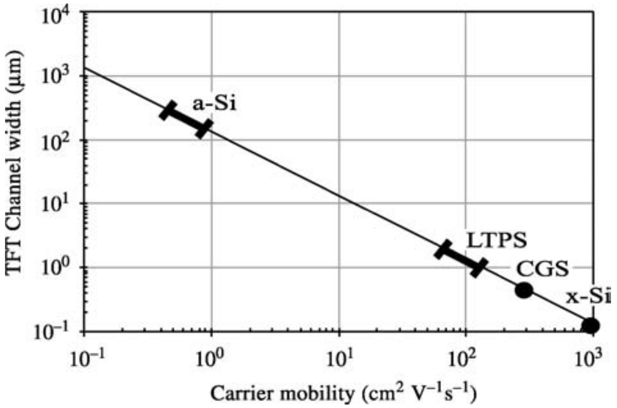


Figure C2.5.12. Channel widths of transistors capable of driving an OLED pixel at 300 cd m^{-2}, assuming an efficiency of 10 cd A^{-1}, pixel pitch of 300 μm and typical values for TFT electrical characteristics.

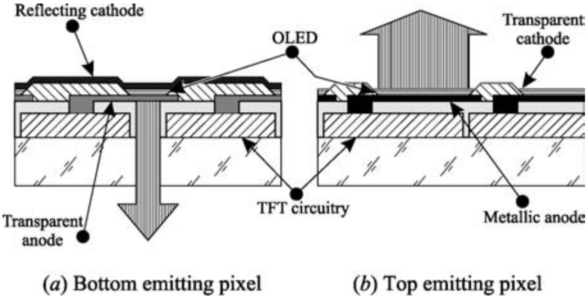

Reflecting cathode
OLED
Transparent cathode
Transparent anode
TFT circuitry
Metallic anode
(a) Bottom emitting pixel
(b) Top emitting pixel

Figure C2.5.13. Cross-sections of bottom-emitting and top-emitting devices.

there is another option that uses a metallic anode and thin transparent cathode, typically referred to as a top emitter (see figure C2.5.13(*b*)). This increase in aperture ratio gives a boost to display lifetime as the current density on an OLED device will be lower, and can also increase efficiency primarily by allowing larger TFT areas (less TFT voltage drop). The disadvantages of this are, at present, significant in that the encapsulation must be both transparent and possess better barrier properties than in the standard case as the thin, highly reactive, electron injector is no longer protected by a thick metal layer.

C2.5.10.1 Voltage programming

Figure C2.5.14(*a*) shows a schematic of a typical AMOLED display panel with figure C2.5.14(*b*) showing the simplest OLED pixel drive circuit. With an LCD all that is required is a method to fix a voltage over a capacitor, so the only transistor required is a select device acting as a switch to connect the cell to the data line when the row on which the pixel resides is addressed. In the case of an OLED

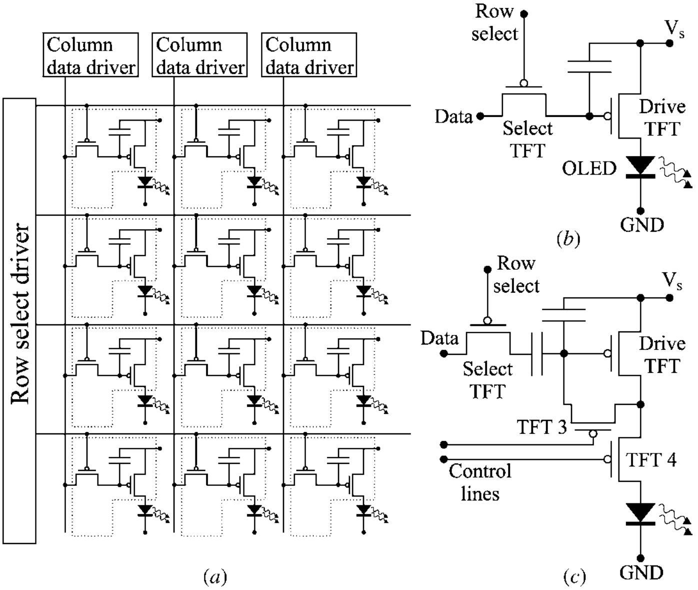

Column data driver
Column data driver
Column data driver
Row select driver
(a)
Row select
Vs
Data
Select TFT
Drive TFT
OLED
GND
(b)
Row select
Vs
Data
Select TFT
Drive TFT
TFT 3
TFT 4
Control lines
GND
(c)

Figure C2.5.14. Voltage-programmed active matrix pixel circuits: (*a*) schematic of an active matrix; (*b*) simplest OLED pixel circuit; (*c*) threshold-voltage compensated circuit.

a drive transistor is needed through which current can be controlled, thus the simplest circuit uses a select TFT to set a gate voltage on a drive transistor, and the gate voltage is set to produce the required luminous output [14]. However, the current controlled through the drive transistor is dependent not only on the gate voltage but also all the TFT characteristics including, in particular, the threshold voltage.

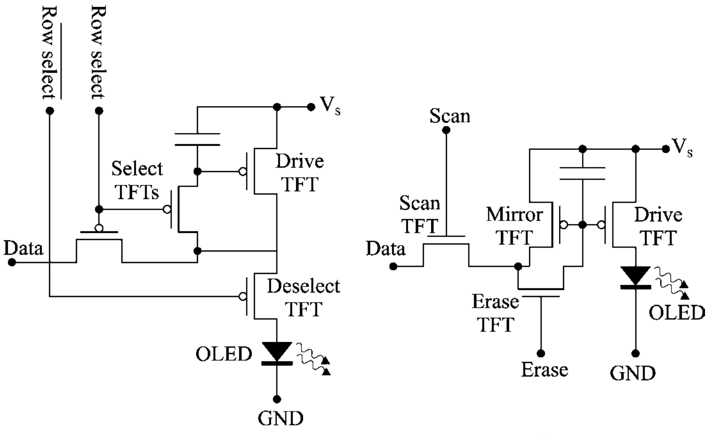
The alternative way of driving the 2-TFT cell is to turn on the transistor hard and use display sub-frames to achieve grey scale. This method does work well; however, it is essentially a voltage-driven method and so suffers from sensitivity to OLED variations and accelerated ageing (compared to current drive).

Attempts have been made to compensate for threshold voltage; one such [15] can be seen in figure C2.5.14(*c*). During the line scan the select transistor is first activated and 0 V applied to the data line. TFT3 is closed which discharges the charge on the gate of the drive TFT down to threshold. TFT4 is then opened, disconnecting the OLED, and TFT3 is opened with the threshold voltage held at the drive TFT gate. The signal applied on the data line will now produce a voltage at the drive TFT gate offset by the threshold voltage. The select transistor is deactivated holding the drive level. This circuit, however, still depends on the uniformity of the other parameters of the drive TFT, as well as requiring two TFTs that can handle the full drive current and two extra control lines. Both of the above circuits also suffer, when analogue driven, from the highly non-linear relationship between gate voltage and drive current.

C2.5.10.2 Current programming

All current drive schemes have the advantages of linearity and relative insensitivity to transistor and OLED variations. Two general methods are possible in principle—the setting of OLED current directly or indirectly through a current mirror.

Figure C2.5.15(*a*) shows a typical example of the first of these circuits of which there are many variations [16, 17]. At the start of the addressing period, the deselect TFT is opened and the select TFTs are closed, diverting the output from the drive TFT to the data line. The display controller sinks a current through the data line and any mismatch between this current and that supplied by the drive TFT will modify the charge on the capacitor until the currents are balanced. At the end of the addressing period, the select TFTs are opened and the deselect TFT is closed, holding the charge on the capacitor and redirecting the drive current to the OLED. This pixel circuit has excellent uniformity and, by its



(*a*) (*b*)

Figure C2.5.15. Current-programmed pixel circuits: (*a*) current sampling circuit; (*b*) amplifying current mirror.

nature, extremely good linearity resulting in a high fidelity of grey scale. Its primary disadvantages are that it requires three TFTs that must be capable of handling the full drive current as well as needing two complementary control lines.

Figure C2.5.15(*b*) shows the current mirror circuit used in a 13 in OLED demonstrator [18]. The operation of this circuit relies upon an amplifying current mirror. An exact relationship between the geometries of the drive and mirror TFTs results in a constant scaling factor between the drain–source current of each when a common voltage is applied to the gate. This relationship can only work when other TFT parameters are sufficiently similar between the two devices. However, this is the case when the TFTs have similar layouts and are in close proximity. During the addressing period, the erase TFT is initially closed to clear the charge on the capacitor. The scan TFT is then closed, a current is sunk through the data line and the voltage on the common gate capacitor adjusts so that the mirror TFT supplies this current. The erase and scan TFTs are then opened and the voltage is held on the gate, producing a scaled current in the drive TFT. This circuit possesses the advantages of the previous circuit without the problems of two control lines or three large TFTs.

Use of either of these pixel cells will produce displays with excellent uniformity and a high fidelity of grey-scale reproduction; however, there are yet some problems remaining, which they do not solve, relating to the OLED devices themselves—burn-in and pixel ageing. While neither of these problems are directly associated with the TFT cell, and might be solved elsewhere, there are pixel circuits which can minimize the effects of these factors.

C2.5.10.3 Optical feedback

An optical feedback TFT cell uses the photocurrent generated by a photodiode, detecting a portion of the OLED light emission, to set or control the gate voltage of the drive TFT. There has been less (public) activity on this type of pixel circuit; however, two potential schemes have emerged.

Figure C2.5.16(*a*) shows a pixel cell presented recently [19]. When the select TFT is closed the data line sets a charge on the gate capacitor. Once the select TFT is opened, any light emitted by the OLED will cause a photocurrent to be generated by the photodiode, gradually discharging the capacitor. If the capacitor has been discharged sufficiently to turn off the drive TFT before the next address period, then the light emitted by the pixel will be proportional to the charge on the capacitor and the response of the photodiode, but, in principle, independent of any TFT or OLED characteristics, including ageing effects.

The pulsed output produced by this pixel cell can cause a number of difficulties. First of all, as it is quite a sharp pulse, a relatively high voltage power rail is required which results in an overall reduction

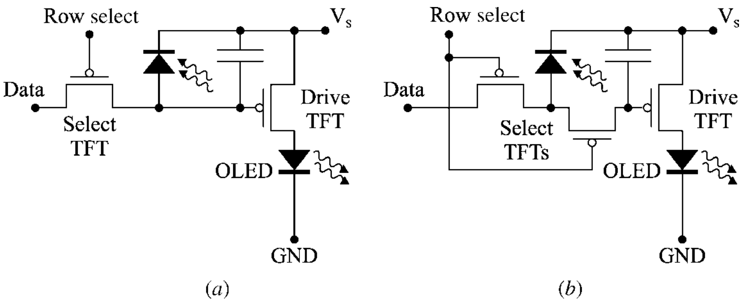


Figure C2.5.16. Optical feedback pixel circuits: (*a*) voltage programmed with discharging photodiode; (*b*) current programmed optical feedback.

in efficiency. Secondly, the higher peak currents may also require a larger drive TFT. Finally, depending on the OLED material, the device lifetime might be adversely affected by the pulsed operation, although of course this cell does compensate for ageing effects.

The last pixel cell, shown in figure C2.5.16(*b*), is a circuit developed to overcome the above difficulties. Its operation is similar to that of the current mirror circuit above, and as such it is sometimes referred to as a current programmed optical feedback. At the start of the line scan the select TFTs are closed, connecting both the data line and photodiode to the drive TFT gate capacitor. A current is sunk through the data line and any mismatch between the photocurrent generated by the photodiode and the requested current will adjust the drive TFT gate voltage (and therefore the OLED drive) until the light generated by the OLED matches that requested on the data line. Once equilibrium is reached, the select TFTs are opened, isolating the gate capacitor from both the data line and the photodiode.

The optical feedback is only operational during the addressing period, thus reducing the circuit's sensitivity to ambient light. Further to this it is possible to fully compensate for ambient light, and light generated by neighbouring pixels, by initially setting 0 V on the data line, detecting the photocurrent generated with the OLED turned off and then using this to offset the signal current.

C2.5.11 Optical display performance

The basic optical performance of any emissive display is possibly the simplest aspect of the display design to consider (unlike LCD displays). As far as an observer of a display is concerned, an OLED is a nearly perfect Lambertian emitter, i.e. no apparent variations in visual appearance with angle. However, as was alluded to in the section on electro-optic response, not much of the generated light is actually emitted from the display. At the time of writing none of the proposed methods to enhance the optical output have made it out of the laboratory; however, the most promising candidates will be reviewed at the end of this section.

C2.5.11.1 Light emission

The optical structure of an OLED can be simplified to a plane of radiating dipoles, in a high-refractive-index medium, a set distance from a metallic mirror (typically the cathode). Depending on the technology, the dipoles are either randomly orientated (small molecule) or have their axes aligned to the plane of the device (polymer). Dipoles preferentially emit radially (perpendicular to their axis) resulting in a uniform emission for the random orientation and an increased emission normal to the device plane for the aligned dipoles. The propagation of the emitted light through the device layers and the substrate to air results, for either case, in a nearly Lambertian emission profile; however, for the polymer a higher fraction of the light is output coupled.

The close proximity of the metallic mirror (<100 nm) complicates matters. The reflected image if the emitting dipoles optically interferes with their originals, modifying the emission profile both in terms of intensity and colour. Weak interference effects from other optical interfaces within the device can also have a lesser effect on the light emission. Typically a device is engineered to keep the dipoles on or close to the optimum location to increase optical output; however, the effect will impose a tolerance on the thicknesses of the organic layers. There is also an observable colour change with angle; however, this effect is usually small.

C2.5.11.2 Contrast

The most significant problem with the reflecting electrode is the reflection of ambient light reducing the contrast of a display. This can be dealt with in one of a number of ways, all of which impair some other aspect of the optical performance. A circular polarizer might typically have a single-pass transmission of 35–40% and a double-pass transmission of only 1–2%. Thus, the reflected light is attenuated by $>98\%$ at the expense of a 60–65% reduction in optical output, which will also become linearly polarized. The Black Layer from Luxell includes interference layers in the cathode structure to make the cathode black, which also attenuates the optical output by 50–60% and reduces the reflection to ~2% without polarization, although this can increase the tolerances required on the other device layers. Finally, louvred privacy filters can help, as the worst reflected light is typically off-axis, but at the expense of viewing angle.

Circular polarizers are the standard method used to increase contrast as they are commonly available in many varieties, relatively inexpensive and simple to apply. The Black Layer is a relatively new technology to OLED displays and, as it needs to be integrated into the display manufacturing line, has yet to appear in production. The applications for which OLED displays have been demonstrated have yet to include those that commonly use louvred filters.

C2.5.11.3 Enhancements

There is a whole range of technologies proposed to increase, or otherwise modify, the light output from OLED displays. To date, only one has been demonstrated in a display (micro-cavity) and that only to provide some colour tuning [18]. All of these techniques either attempt to directly modify the light emission process, or re-format the emission at the viewing side of the substrate. A detailed review of each is beyond the scope of this paper; however, many reviews on the subject exist in the literature (for example [20]).

C2.5.12 Summary

In the 15 years since invention, OLEDs have developed from a barely visible laboratory curiosity to introductory commercial products. They still have a long way to go to fulfil their potential and many organizations are now devoting significant effort towards this.

By going step by step through the fabrication process and the issues related to driving for a variety of display architectures, we hope to have shown the reader where the strengths and weaknesses of the technology lie, both for organics in general and the differences between polymers and small-molecule-based devices.

Organic devices have the potential to make bright, attractive and colourful flat panel emissive displays. They have the fundamental advantages over other display technologies of both a simple display structure and low-voltage operation coupled with an intrinsically wide viewing angle and low-power operation.

Small-molecule devices currently have the commercial lead with several products already available. However, they have a significant hurdle to cross to get to large displays due to the process difficulties associated with evaporation through large-area high-resolution shadow masks.

Polymer devices need to prove that ink-jet printing is a viable manufacturing technology in order to get to full colour, but once this is done large displays should follow.

References

[1] Tang C W and Van Slyke S A 1987 Organic electroluminescent diodes *Appl. Phys. Lett.* **51** 913–915

[2] Burroughes J H, Bradley D D C, Brown A R, Marks R N, MacKay K, Friend R H, Burn P L and Holmes A B 1990 Light-emitting diodes base on conjugated polymers *Nature* **347** 539–541

[3] Partridge R H 1983 Electroluminescence from polyvinylcarbazole films *Polymer* **24** 733–762

[4] Helfrich W and Schneidere W G 1965 *Phys. Rev. Lett.* **14** 229

[5] Wilson J S, Dhoot A S, Seeley A J A B, Khan M S, Köhler and Friend R H 2001 Spin-dependent exciton formation in π-conjugated compounds *Nature* **413** 828–831

[6] Groenendaal L, Jonas F, Freitag D, Pielartzik H and Reynolds J R 2000 Poly(3,4-ethylenedioxythiophene) and its derivatives: past present and future *Adv. Mater.* **12** 481–494

[7] Burrows P E *et al* 2001 Gas permeation and lifetime tests on polymer OLEDs *Proc. SPIE* **4105** 75–83

[8] Brown T M, Friend R H, Millard I S, Lacey D J, Burroughes J H and Cacialli F 2001 Efficient electron injection in blue-emitting polymer light-emitting diodes with LiF/Ca/Al cathodes *Appl. Phys. Lett.* **79** 174–176

[9] Duinevald P C, de Kok M M, Buechel M, Sempel A H, Mutsaers K A H, van de Weijer P, Camps I G J, van den Biggelaar T J M, Rubingh J J M and Haskel E I 2002 Ink-jet printing of polymer light-emitting devices *Proc. SPIE* **4464** 59–67

[10] McElvain J, Antoniadis H, Hueschen M R, Miller J N, Roitman D M, Sheats J R and Moon R L 1996 Formation and growth of black spots in organic light-emitting diodes *J. Appl. Phys.* **80** 6002–6007

[11] Kim J-S, Ho P K H, Murphy C E, Baynes N and Friend R H 2001 Nature of non-emissive black spots in polymer light-emitting diodes by in-situ micro-Raman spectroscopy *Adv. Mater.* **14** 206–209

[12] Murgatroyd P N 1970 *J. Phys. D* **3** 151

[13] Kim J-S, Ho P K H, Greenham N C and Friend R H 2000 Electroluminescence emission pattern of organic light-emitting diodes: implications for device efficiency calculations *J. Appl. Phys.* **88** 131

[14] Brody T P, Luo F C, Szepsi Z P and Davies D H 1975 *IEEE Trans Electron. Dev.* **ED-22** 739–748

[15] Dawson R *et al* 1998 The impact of the transient response of organic light-emitting diodes on design of active matrix OLED displays *IEEE International Electron Device Meeting* 875

[16] Hunter I M, Young N D, Johnson M T and Young E W A 1999 *SID Proceedings of the Sixth International Display Workshop* pp 1095–1096

[17] He Y, Hattori R and Kanicki J 2000 *IEEE Electron. Dev. Lett.* **21** 590

[18] Sasaoka T *et al* 2001 *SID Int. Symp. Digest* **32** 384–386

[19] Fish D, Young N, Childs M, Steer W, George D, McCulloch D, Godfrey S, Trainer M, Johnson M, Giraldo A, Lifka H and Hunter I 2002 *SID Int. Symp. Digest* **33** 968–971

[20] Patel N K, Cinà S and Burroughes J H 2002 *IEEE J. Sel. Top. Quantum Electron.* **8** (2)

Further reading

Bulovic V, Burrows P E and Forrest S R 1999 Molecular organic light-emitting devices *Semiconductors and Semimetals* vol 64, ed G Mueller (New York: Academic) pp 255–306

Bulovic V and Forrest S R 2000 Polymeric and molecular organic light-emitting devices: a comparison *Semiconductors and Semimetals* vol 65, ed G Mueller, R K Willardson and E R Weber (New York: Academic) pp 1–26

Hartnagel H L, Dawar A L, Jain A K and Jagadish C 1995 *Semiconducting Transparent Thin Films* (Bristol: Institute of Physics Publishing)

Sato Y 1999 Organic LED system considerations *Semiconductors and Semimetals* vol 64, ed G Mueller (New York: Academic) pp 209–254

Stanford Resources 1998 *Flat Information Displays Market and Technology Trends* (Stanford Resources)

Cambridge Display Technology: www.cdtltd.co.uk

Kodak Display Products: www.kodak.com/US/en/corp/display/index.jhtml

Universal Display Corporation: www.universaldisplay.com

Dow light emitting polymers: www.dow.com/pled/

Covion light emitting polymers: www.covion.com

Idemitsu Kosan: www.idemitsu.co.jp/e/

C2.6
Three-dimensional display systems

Nick Holliman

C2.6.1 Introduction

Today's three-dimensional display systems provide new advantages to end-users; they are able to support an auto-stereoscopic, no-glasses, three-dimensional experience with significantly enhanced image quality over previous generation technology. There have been particularly rapid advances in personal auto-stereoscopic three-dimensional display for desktop users brought about because of the opportunity to combine micro-optics and LCD displays coinciding with the availability of low-cost desktop image processing and three-dimensional computer graphics systems.

In this chapter, we concentrate our detailed technical discussion on personal three-dimensional displays designed for desktop use as these are particularly benefiting from new micro-optic elements. We emphasize the systems aspect of three-dimensional display design believing it is important to combine good optical design and engineering with the correct digital imaging technologies to obtain a high-quality three-dimensional effect for end-users. The general principles discussed will be applicable to the design of all types of stereoscopic three-dimensional display.

C2.6.2 Human depth perception

Defining the requirements for three-dimensional display hardware and the images shown on them is an important first step towards building a high-quality three-dimensional display system. We need a clear understanding of how a digital stereoscopic image is perceived by an end-user in order to undertake valid optimization during the design process.

Binocular vision provides humans with the advantage of depth perception derived from the small differences in the location of homologous, or corresponding, points in the two images incident on the retina of the eyes. This is known as stereopsis (literally solid seeing) and can provide precise information on the depth relationships of objects in a scene.

The human visual system also makes use of other depth cues to help interpret the two images incident on the retina and from these build a mental model of the three-dimensional world. These include monocular depth cues (also known as pictorial [18] or empirical [39] cues), whose significance is learnt over time, and oculomotor cues in addition to the stereoscopic cue [39]. We consider these in turn and introduce in detail binocular vision both in the natural world and when looking at an electronic three-dimensional display.

C2.6.2.1 Monocular and oculomotor depth cues

Redundancy is built into the visual system and even people with monocular vision are able to perform well when judging depth in the real world. Therefore, in the design of three-dimensional displays, it is important to be aware of the major contribution of monocular two-dimensional depth cues in depth perception and aim to provide displays with at least as good basic imaging performance as two-dimensional displays. Ezra [12] suggests this should include levels of brightness, contrast, resolution and viewing range that match a standard two-dimensional display with the addition of the stereoscopic cue provided by generating a separate image for each eye.

The monocular depth cues are experiential and over time observers learn the physical significance of different retinal images and their relation to objects in the real world. These include:

- Interposition: objects occluding each other suggest their depth ordering.
- Linear perspective: the same size object at different distances projects a different size image onto the retina.
- Light and shade: the way light reflects from objects provides cues to their depth relationships; shadows are particularly important in this respect.
- Relative size: an object with smaller retinal image is judged further away than the same object with a larger retinal image.
- Texture gradient: a texture of constant size objects, such as pebbles or grass, will vary in size on the retina with distance.
- Aerial perspective: the atmosphere affects light travelling through it, for example due to fog, dust or rain. As light travels long distances, it is scattered, colours lose saturation, sharp edges are diffused and colour hue is shifted towards blue.

Many of these cues are illustrated in figure C2.6.1 and can be considered to be two-dimensional depth cues because they are found in purely monoscopic images. Two other non-binocular depth cues are available: motion parallax and oculomotor cues.

Motion parallax provides the brain with a powerful cue to three-dimensional spatial relationships without the use of stereopsis [18, 39] and this is the case when either an object in the scene or the observer's head moves. Motion parallax does not, however, make stereopsis redundant, as comprehending images of complex scenes can be difficult without binocular vision. Yeh [66] and others have shown that both stereopsis and motion parallax combined result in better depth perception than either cue alone.

Oculomotor depth cues are due to feedback from the muscles used to control the vergence and accommodation of the eye. They are generally regarded as having limited potential to help depth judgement [16, 39, 41] and we will move on to consider how human binocular vision works when used to view the natural world.

C2.6.2.2 Binocular depth perception in the natural world

Extracting three-dimensional information about the world from the images received by the two eyes is a fundamental problem for the visual system. In many animals, perhaps, the best way of doing this comes from the binocular disparity that results from two forward facing eyes having a slightly different viewpoint of the world [5]. The binocular disparity is processed by the brain giving the sensation of depth known as stereopsis.

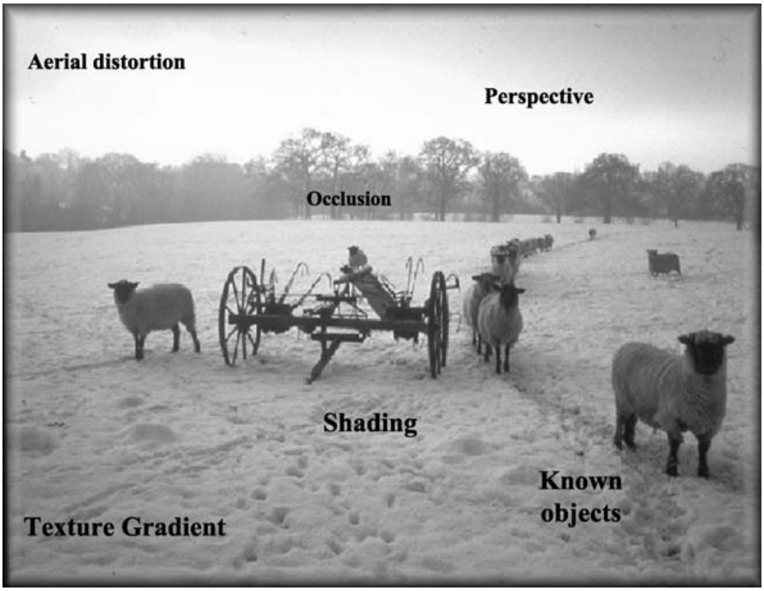

Aerial distortion
Perspective
Occlusion
Shading
Known objects
Texture Gradient

Figure C2.6.1. Picture illustrating the depth cues available in a two-dimensional image (photographer David Burder).

Stereo depth perception in the natural world is illustrated in figure C2.6.2. The two eyes verge the visual axes so as to fixate the point F and adjust their accommodation state so that points in space at and around F come into focus.

The vergence point, F, projects to the same position on each retina and therefore has zero retinal disparity, i.e. there is no difference between its location in the left and right retinal images. Points in front or behind the fixation point project to different positions on the left and right retina and the resulting binocular disparity between the point in the left and right retinal images provides the observer's brain with the stereoscopic depth cue. Depth judgement is therefore relative to the current vergence point, F, and is most useful to make judgements on the relative rather than absolute depth of objects in a scene.

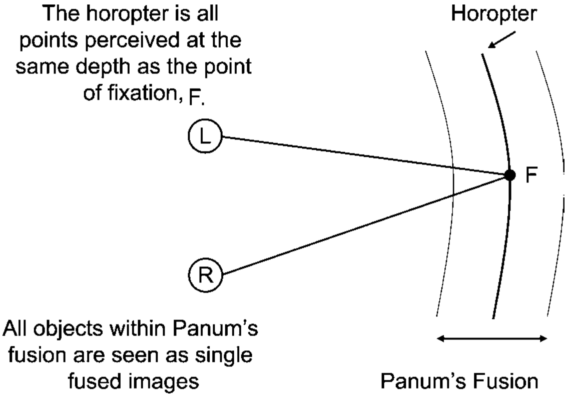

The horopter is all points perceived at the same depth as the point of fixation, F.
Horopter
L
F
R
All objects within Panum's fusion are seen as single fused images
Panum's Fusion

Figure C2.6.2. The geometry of the binocular vision when viewing the natural world.

Points in space, other than F, which project zero retinal disparity are perceived to lie at the same depth as the vergence point; all points that project zero retinal disparity are described as being on a surface in space known as the horopter. The shape of the horopter shown in figure C2.6.2 is illustrative only; it is known in practice to be a complex shape and to have non-linear characteristics [3, 18].

Geometrically, we can define angular disparity, α, as the difference between the vergence angle at the point of fixation, F, and the point of interest. Considering figure C2.6.3:

Points behind the fixation point, such as A, have positive disparity.

$$\alpha_a = f - a. \tag{C2.6.1}$$

Points in front of the fixation point, such as B, have negative disparity.

$$\alpha_b = f - b. \tag{C2.6.2}$$

The smallest perceptible change in angular disparity between two small objects is referred to as stereo acuity, δ [65]. The advantage of defining stereo acuity as an angle is that it can be assumed to be constant regardless of the actual distance to and between the points A and B. However, it is also helpful to know how this translates in terms of the smallest perceived distance between objects at the typical viewing range of a desktop three-dimensional display. This will allow us to compare the ability of the eye to perceive depth with the ability of different displays designs to reproduce it.

Considering figure C2.6.4 when points A and C can just be perceived to be at a different depth, then stereo acuity will be:

$$\delta = a - c. \tag{C2.6.3}$$

Various studies [28, 31, 65] show the eye is able to distinguish very small values of δ, as little as $1.8''$ (seconds of arc). As the exact limits vary between people, Diner and Fender [8] suggest that a practical working limit is to use a value of stereo acuity $\delta = 20''$. Using this value, we can calculate the size of the smallest distinguishable depth difference at a given distance from the observer. We choose $m = 750$ mm as the distance from the observer as a common viewing distance for desktop stereoscopic displays and use an average eye separation, $e = 65$ mm.

Calculating along the centre line between the visual axes, we can find the minimum distinguishable depth, n, at distance m by considering points A and C. The angle a can be calculated as:

$$a = 2 \times \arctan\left(\frac{e/2}{m}\right) = 2 \times \arctan\left(\frac{32.5}{750}\right) \tag{C2.6.4}$$

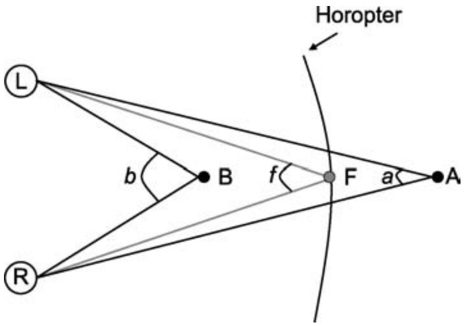


Figure C2.6.3. Angular disparity is defined relative to the current fixation point.

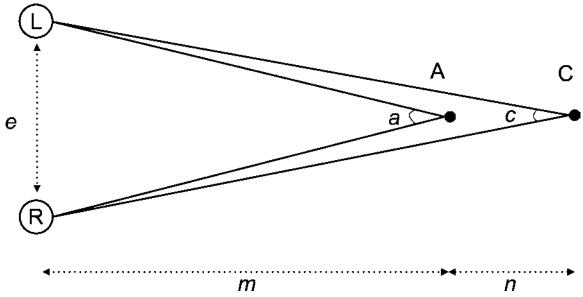


Figure C2.6.4. Stereo acuity defines smallest depth difference an observer can perceive.

by the definition of stereo acuity we know that:

$$\tan(c/2) = \tan\left(\frac{a-\delta}{2}\right) = \tan\left(\frac{a-20''}{2}\right) \tag{C2.6.5}$$

and if n is the distance between A and C we also know that:

$$\tan(c/2) = \frac{e/2}{m+n} \tag{C2.6.6}$$

rearranging equation (C2.6.6) we have:

$$n = \left(\frac{e/2}{\tan(c/2)}\right) - m. \tag{C2.6.7}$$

Substituting equation (C2.6.4) in equation (C2.6.5) and using the result to solve equation (C2.6.7) gives $n = 0.84\,\text{mm}$.

We can conclude that a person with a stereo acuity of $20''$ and an eye separation of 65 mm will be able to perceive depth differences between small objects of just 0.84 mm at a distance of 750 mm from the eyes.

It is also possible to calculate a geometric value for the furthest possible range of stereo vision which occurs when the vergence angle between the two visual axes is equal to or less than the stereo acuity.

The distance m from the observer to the point A when the angle $a = \delta$ is given by

$$m = \frac{e/2}{\tan(a/2)}. \tag{C2.6.8}$$

Again taking $\delta = 20''$ and $e = 65\,\text{mm}$ we get $m = 670\,\text{m}$.

This means that points such as C at a distance of 670 m or more from the observer will not be able to be distinguished in distance from A using binocular vision alone. Just before this limit is reached, the smallest distinguishable depth difference between points will have increased to over 300 m and it is clear only gross differences in depth will be perceived at the furthest limits of stereoscopic perception.

To summarize the above, binocular vision uses the stereoscopic depth cue of retinal disparity to perceive an object's depth relative to the fixation point of the two eyes. At close and near range this provides a high degree of depth discrimination and even at tens of metres from the observer enables relative depth perception for larger objects.

C2.6.2.3 Depth perception in electronic stereoscopic images

Wheatstone [60] demonstrated that the stereoscopic depth sensation could be recreated by showing each eye a separate two-dimensional image. The left and right eye views should be two-dimensional planar images of the same scene from slightly different viewpoints; the difference in the viewpoints generates disparity in the images. When the images are subsequently viewed, the observer perceives depth in the scene because the image disparity creates a retinal disparity similar, but not identical, to that seen when looking directly at a natural scene.

Wheatstone was able to demonstrate this effect by building the first stereoscope and many devices have since been invented for stereoscopic image presentation each with their own optical configurations. Reviews of these devices and the history of stereoscopic imaging are available in several sources [23, 30, 32, 40, 53].

To help characterize and compare the performance of different electronic three-dimensional display designs, we will consider the perception of depth in planar stereo image pairs and how this differs from the stereoscopic perception of depth in the natural world.

A key physiological difference is that although the eyes need to verge off the stereoscopic image plane to fixate points in depth, their accommodation state must always keep the image plane itself in focus. This requires the observer to be able to alter the normal link between vergence and accommodation and is one reason why images with large perceived depth are hard to view. This suggests that the perceived depth range in stereoscopic image pairs needs to be limited to ensure the observer will find a stereo image pair comfortable to view.

While there are several studies of the comfortable perceived depth range on electronic three-dimensional displays [17, 64, 65], it can be difficult to factor out variables relating to display performance from the results. Display variables include absolute values, and inter-channel variations, of brightness and contrast in addition to stereoscopic image alignment and crosstalk. All of these can affect the comfortable range of perceived depth on a particular display. For example, high-crosstalk displays generally do not support deep images as the ghosting effect becomes more intrusive to the observer as screen disparity is increased.

An analysis of the geometry of perceived depth assuming a display with ideal properties helps identify the geometric variables affecting perceived depth independently of the display used. Geometric models of perceived depth have been studied by Helmholtz [23] and Valyus [53] and more recently in [8, 24, 27, 64]. We present a simplified model in figure C2.6.5 for discussion purposes which helps emphasize the key geometric variables affecting the perception of stereoscopic images.

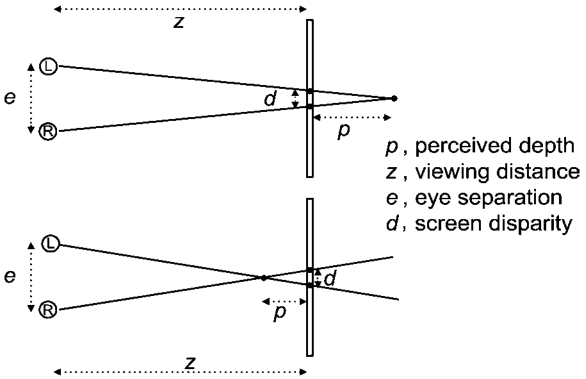


Figure C2.6.5. Perceived depth behind (top) and in front (bottom) of the display plane.

Figure C2.6.5 shows the geometry of perceived depth for a planar stereoscopic display; for simplicity, we consider the geometry along the centre line of the display only; more general expressions are available [23, 64]. The viewer's eyes, L and R, are separated by the interoccular distance, e, and are at a viewing distance, z, from the display plane. The screen disparity between corresponding points in the left and right images, d, is a physical distance resulting from the image disparity which is a logical value measured in pixels. Image disparity is constant for a given stereo pair; however, screen disparity will vary depending on the characteristics of the physical display. Screen disparity in a pair of aligned stereo images is simply the difference of the physical x coordinates of corresponding points in the right x_r and left x_l images:

$$d = x_\mathrm{r} - x_\mathrm{l} \tag{C2.6.9}$$

Two key expressions relating screen disparity to perceived depth can be derived from the similar triangles in figure C2.6.5. Perceived depth behind the screen plane, i.e. positive values of d, is given by:

$$p = \frac{z}{\left(\frac{e}{|d|}\right) - 1} \tag{C2.6.10}$$

Perceived depth in front of the screen plane, i.e. negative values of d, is given by:

$$p = \frac{z}{\left(\frac{e}{|d|}\right) + 1} \tag{C2.6.11}$$

Equations (C2.6.10) and (C2.6.11) provide several insights into the geometric factors affecting perceived depth:

- z, the viewing distance to the display. Perceived depth is directly proportional to the viewing distance, z. Therefore a viewer looking at the same stereoscopic image from different distances perceives different depth. How important this is, is application dependent, but applications such as CAD, medical imaging and scientific imaging may critically depend on accurate and consistent depth judgements.
- d, the screen disparity. Perceived depth is also directly proportional to screen disparity, d. The screen disparity for any given stereoscopic image varies if the image is displayed at different sizes, either in different size windows on the same screen or on different size screens. Again this is important to note in applications where depth judgement is a critical factor. It means stereoscopic images are display dependent and an image displayed on a larger display than originally intended could exceed comfortable perceived depth limits or give a false impression of depth.
- e, individual eye separation. Perceived depth is inversely proportional to individual eye separation which varies over a range of approximately 55–75 mm with an average value often taken as 65 mm. Children can have smaller values of eye separation and therefore see significantly more perceived depth in a stereoscopic image than the average adult. It may be particularly important to control perceived depth in systems intended for use by children, as they will reach the limits of their vergence/accommodation capabilities sooner than most adults.

For display design, controlling these variables so that the viewer sees a consistent representation of depth ideally requires tracking head position, identifying eye separation and controlling screen disparity. These are challenging goals in addition to designing a display with imaging performance as good as a two-dimensional display.

C2.6.2.4 Benefits of binocular vision

An important question is what advantages does binocular vision provide in the real world? As a visual effect, it clearly fascinates the majority of people when they see a three-dimensional picture. Beyond the attractive nature of stereoscopic three-dimensional images, they provide the following benefits over monocular vision:

- Relative depth judgement. The spatial relationship of objects in depth from the viewer can be judged directly using binocular vision.
- Spatial localization. The brain is able to concentrate on objects placed at a certain depth and ignore those at other depths using binocular vision.
- Breaking camouflage. The ability to pick out camouflaged objects in a scene is probably one of the key evolutionary reasons for having binocular vision [47].
- Surface material perception. For example, lustre [23], sparkling gems and glittering metals are in part seen as such because of the different specular reflections detected by the left and right eyes.
- Judgement of surface curvature. Evidence suggests that curved surfaces can be interpreted more effectively with binocular vision.

These benefits make stereo image display of considerable benefit in certain professional applications where depth judgement is important to achieving successful results. In addition, the effect of stereopsis is compelling enough that stereoscopic images have formed the basis of many entertainment systems.

C2.6.3 Three-dimensional display designs using micro-optics

The possible combinations of LCD and micro-optics provide many degrees of freedom for display design; the ideal three-dimensional display design will depend on specific application requirements. However, there are characteristics that all display designs should give consideration to and we briefly review these here.

There is a need to compare the basic image quality of a three-dimensional design to that achieved by current two-dimensional displays; i.e. the two-dimensional characteristics of a three-dimensional display should match the performance of two-dimensional displays as closely as possible. Key characteristics are:

- brightness typical of a current LCD display is $150\,\mathrm{cd\,m^{-2}}$;
- contrast typical of a current LCD display is 300:1;
- colour reproduction, measured white points and measured CIE coordinates of primaries.

These values are typical of current two-dimensional displays but are clearly a moving target as two-dimensional displays improve.

In addition, there are a number of important characteristics unique to three-dimensional displays. The first is that two-dimensional characteristics need to be matched between all the viewing windows of the three-dimensional display. Each viewing window should also be matched spatially and temporally so that there is no noticeable position or time difference between corresponding images.

Inter-channel crosstalk appears to an observer as a ghost image, which will be particularly visible at high-contrast edges in images. It is an unwanted feature in most display designs because high values of crosstalk are known to be detrimental to three-dimensional effect, particularly on high-contrast displays showing large values of perceived depth [42]. Ideally crosstalk measurements need to be no more than 0.3% if the ghosting effect is to be imperceptible to an observer. Crosstalk, although often due to optical effects in the display, can also result from poor separation of the two image channels in the display driving electronics, image compression formats or the camera system generating the images.

An observer of a two-dimensional display will usually expect to be able to see a good quality image at a wide range of positions in front of the display. Because of the need to direct images separately to the two eyes, many three-dimensional displays have a more limited viewing freedom. Consideration needs to be given to the targets for lateral, vertical and perpendicular freedom in a display design. Three-dimensional display systems capable of supporting multiple observers will often do so at the expense of viewing freedom. Improved viewing freedom can be found in designs with multiple viewing windows or using head tracking to steer viewing windows to follow the observers' head movements. When head tracking is used, a design needs to consider targets for the maximum supported head speed as this directly determines key tolerances.

Some displays have the capability to operate in either three-dimensional or two-dimensional modes switching electronically or mechanically between the two. In this case, the image quality in each mode needs to be considered against the performance of a standard two-dimensional display, as a display in three-dimensional mode will often have different optical performance to the same display in two-dimensional mode.

The capability of a three-dimensional display to represent perceived depth is probably the single most important design target; however, we will return to how to quantify and compare this between displays after presenting details of representative three-dimensional display designs.

We would like three-dimensional displays to provide the ability for the observer to accommodate naturally at the fixation point. However, this is not a feature supported in stereoscopic images and has been attempted in very few display designs.

C2.6.3.1 Stereoscopic systems

Stereoscopic displays require users to wear a device, such as analysing glasses, that ensures left and right views are seen by the correct eye. Many stereoscopic display designs have been proposed and there are reviews of these in several sources [30, 32, 34, 40, 53]. Most of these are mature systems and have become established in several professional markets but suffer from the drawback that the viewer has to wear, or be very close to, some device to separate the left and right eye views. This has limited the widespread appeal of stereoscopic systems as personal displays for home and office use even when the three-dimensional effect is appealing. However, stereoscopic displays are particularly suited to multiple observer applications such as cinema and group presentation where directing individual images to each observer becomes difficult compared to providing each observer with a pair of analysing glasses.

As stereoscopic display systems are well described elsewhere, we limit ourselves here to a summary of the major types using electronic displays:

- Wheatstone mirror stereoscopes using CRT displays or LCD displays.
- Polarized glasses in combination with a method of polarizing the two views.

- Shutter glasses working in synchronization with a view switching display.
- Analglyph glasses analysing different colour channels to obtain the images.
- Brewster stereoscopes, of which head mounted displays are up to date examples.

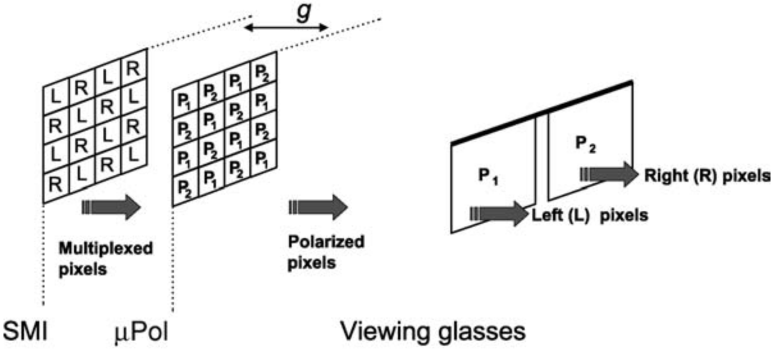
A series of stereoscopic display designs that use polarizing micro-optics have been produced [14], as shown in figure C2.6.6. The micro-optics split a single display into two differently polarized views, which are viewed correctly by left and right eyes when the observer wears a pair of analysing polarized glasses. This requires two half resolution views and may be achieved using a chequerboard pattern of image multiplexing and polarization as shown in figure C2.6.6 as the spatially multiplexed image (SMI) and patterned micro-polarizer (μPol).

A drawback of the design, particularly for direct view LCD-based displays, is the parallax between the display pixels and the micro-polarizer when the micro-polarizer is mounted over the LCD due to the layer of substrate between the two elements forming the gap *g* in figure C2.6.6. If the head moves from the nominal viewing position, part of the adjacent view's pixel becomes visible resulting in crosstalk. One way to reduce this is to use interlace the images in alternate rows so at least lateral head movement is not affected by parallax. As noted by Harrold [20] this problem can only be fully solved in the long term by manufacturing the micro-polarizer element within the LCD pixel cells reducing the parallax between polarizer and pixel.

C2.6.3.2 Auto-stereoscopic systems

Auto-stereoscopic displays are those that do not require the observer to wear any device to separate the left and right views and instead send them directly to the correct eye [6]. This removes a key barrier to acceptance of three-dimensional displays for everyday use but requires a significant change in approach to three-dimensional display design. Auto-stereoscopic displays using micro-optics in combination with an LCD element have become attractive to display designers and several new three-dimensional display

A spatially multiplexed image (SMI) with left (L) and right (R) image pixels is placed behind a patterned micro-polariser (uPol) element.



When viewed with polarized glasses the P1 polarized pixels are seen only in the left eye and P2 polarized in the right.

Figure C2.6.6. The micro-polarizer stereoscopic display principle.

types are now available commercially. The key optical reasons [62] for combining micro-optics with LCD elements are:

- LCDs offer pixel position tolerances better than 0.1 μm.
- LCD pixels, unlike CRT pixels, have high positional stability.
- LCD elements have carefully controlled glass thickness.

Auto-stereoscopic displays have been demonstrated using a range of optical elements in combination with an LCD including:

- Parallax barriers, optical apertures aligned with columns of LCD pixels.
- Lenticular optics, cylindrical lenses aligned with columns of LCD pixels.
- Micropolarizers are found in several auto-stereoscopic three-dimensional display designs.
- Holographic elements have been used to create real images of a diffuse light source.

In the following, we introduce how these elements are used in auto-stereoscopic three-dimensional display designs including two-view and multi-view designs. We begin by looking at auto-stereoscopic two-view designs using twin-LCD elements.

C2.6.3.3 Two-view twin-LCD systems

A successful approach to building high-quality auto-stereoscopic displays has been to use two LCD elements and direct the image from one to the left eye and from the other to the right eye; the principle is illustrated in figure C2.6.7. Several designs have adopted this approach including [12, 13, 22].

Ezra [12, 13] describes one of the designs, which produces bright, high-quality, full colour moving three-dimensional images over a wide horizontal viewing range. As shown in figure C2.6.8, the display produces two viewing windows using a single illuminator. The arrangement of optical elements generates horizontally offset images of the illuminator at a nominal viewing distance to form the viewing windows. An observer's eye placed in one of the viewing windows will see an image from just one of the LCD elements.

If a stereo pair of images is placed on the left and right LCD elements, respectively, then an observer will see a stereoscopic three-dimensional image. The image appears in the plane of the left LCD as the

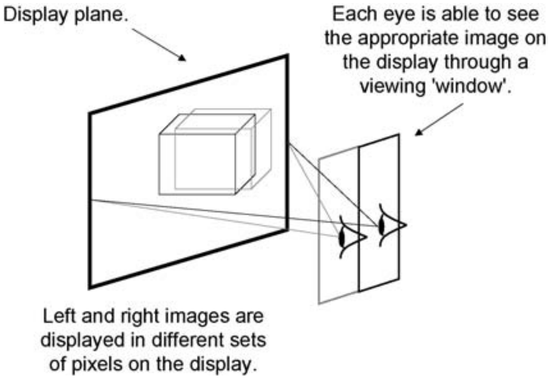

Display plane.
Each eye is able to see the appropriate image on the display through a viewing 'window'.
Left and right images are displayed in different sets of pixels on the display.

Figure C2.6.7. Two-view displays create two viewing windows.

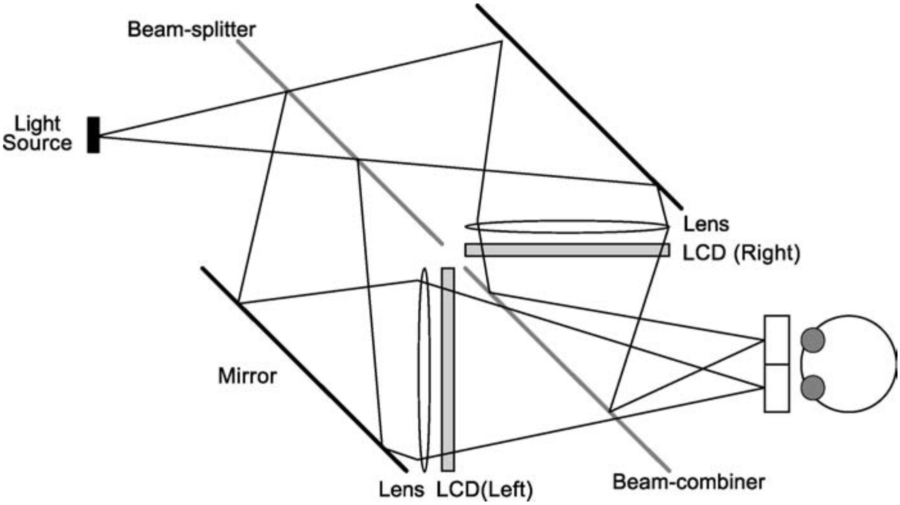


Figure C2.6.8. The twin-LCD display [12].

observer looks at the display and depth is perceived in front and behind this plane. As the two LCDs are seen separately, each eye has a full resolution image and the interface is simply two synchronized channels of digital or analogue video which can be generated at low cost on a desktop PC system.

This basic configuration can be enhanced in several ways: if the light source is moved then the viewing windows can be steered to follow the observer's head position. In order to implement window steering, new technologies for tracking head position have also been developed [25]. The effect of implementing head tracking linked to window steering is to increase the viewing freedom of the display and if the images are updated the design has been demonstrated to provide a full look-around effect. This allows the observer to look around the display and see different views of the scene as they would in the natural world. Image generation for look-around can be implemented by using a three-dimensional computer graphics system to generate the new views when given head tracking position information.

Another possible development [13] is to have multiple light sources providing multiple stereo views to multiple viewers. This could be implemented either by sending the same image pair to each viewer, or by time slicing the light source and the displays to send a different image to each viewer in rapid succession.

The system uses bulk optics and therefore has a large footprint, particularly as the LCD display diagonal size increases. This led to the micro-optic twin-LCD display [61] which provides the same effect in a smaller footprint and is more practical for scaling to larger display sizes.

The micro-optic twin-LCD display is illustrated in figure C2.6.9. The two LCD elements remain in the design with a half mirror acting as a beam combiner between them. The arrangement of optical elements behind one LCD panel directs light so that it forms one viewing window at a nominal viewing distance from the display, another is formed adjacent to this from the backlight of the other LCD panel. As with the bulk optic display the observer placing their eyes in the viewing windows will see the appropriate image in each eye and experience a stereoscopic three-dimensional effect.

As discussed in [61] the micro-optic display produces a better viewing window profile than the bulk-optic display. This is because the micro-optics form a wider and more even illumination distribution for each viewing window so that, when steered, the windows can be moved further laterally before

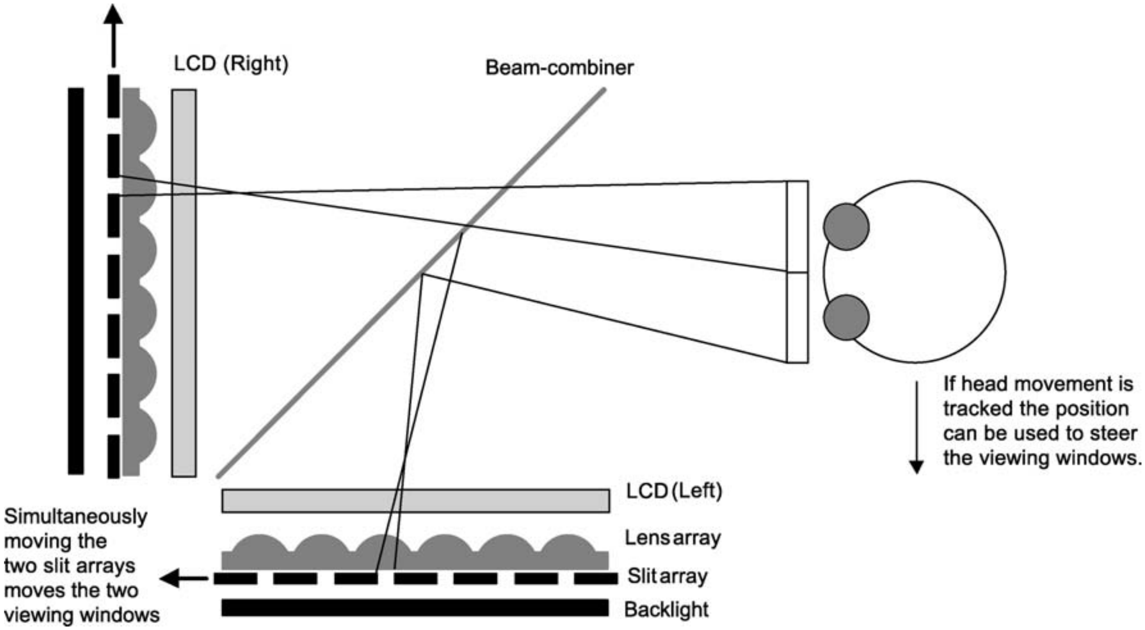

LCD (Right)
Beam-combiner
If head movement is tracked the position can be used to steer the viewing windows.
LCD (Left)
Lens array
Slit array
Backlight
Simultaneously moving the two slit arrays moves the two viewing windows

Figure C2.6.9. The micro-optic twin-LCD display [61].

aberrations reduce their quality. This also results in side lobes of better quality, which in untracked displays can be used by additional observers.

C2.6.3.4 A note on viewing windows

One of the key influences on the perceived performance of auto-stereoscopic displays is the quality of the viewing windows that can be produced at the nominal viewing position. Degradation of the windows due to unresolved issues in the optical design can lead to flickering in the image, reduced viewing freedom and increased inter-channel cross-talk. All of these reduce the quality of viewing experience for observers in comparison to the two-dimensional displays they are used to using. In addition, in head-tracked systems degraded window quality can lead to harder constraints on the accuracy and response speed of the tracking and window steering systems, increasing system costs [25].

The auto-stereoscopic displays considered so far produce two viewing windows in space typically at a nominal distance from the display in a plane parallel to the display surface, as shown in figure C2.6.7. Although often illustrated in two dimensions, the viewing windows have a three-dimensional shape and from above appear as diamonds tapering away from the nominal viewing plane as shown in figure C2.6.10. As long as an observer's pupils stay within these diamonds, and the display is showing a stereo image, the observer will see a three-dimensional image across the whole of the display.

Experimentally the window intensity profile can be determined by measurements using a 1 mm pinhole, a photometric filter and a detector. To fully characterize a display performance, the profile measurements should be repeated at a range of positions vertically and longitudinally offset from the nominal window position. The variables characterizing the quality of the viewing windows are discussed in [61] and are summarized here in figure C2.6.11.

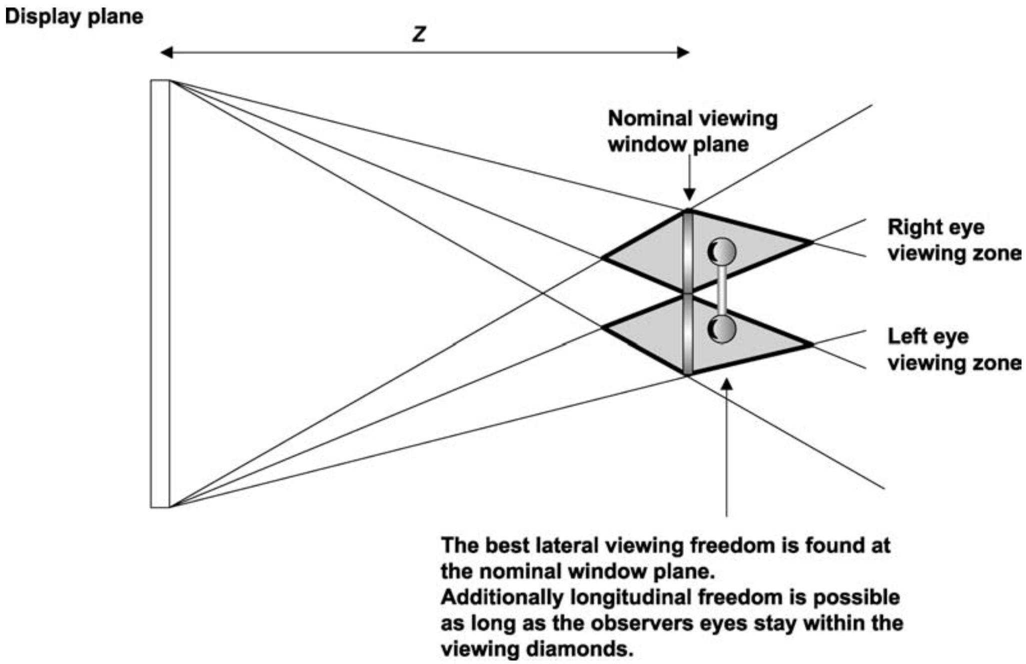

Display plane
Z
Nominal viewing window plane
Right eye viewing zone
Left eye viewing zone
The best lateral viewing freedom is found at the nominal window plane.
Additionally longitudinal freedom is possible as long as the observers eyes stay within the viewing diamonds.

Figure C2.6.10. Viewing freedom in an auto-stereoscopic display [61].

The useful width of the window determines how far an observer can move before the image quality degrades. Larger useful width, up to the interoccular separation, typically 65 mm, provides more comfortable viewing in fixed position displays as there will be a small but useful lateral range of head positions at which a good three-dimensional image can be seen.

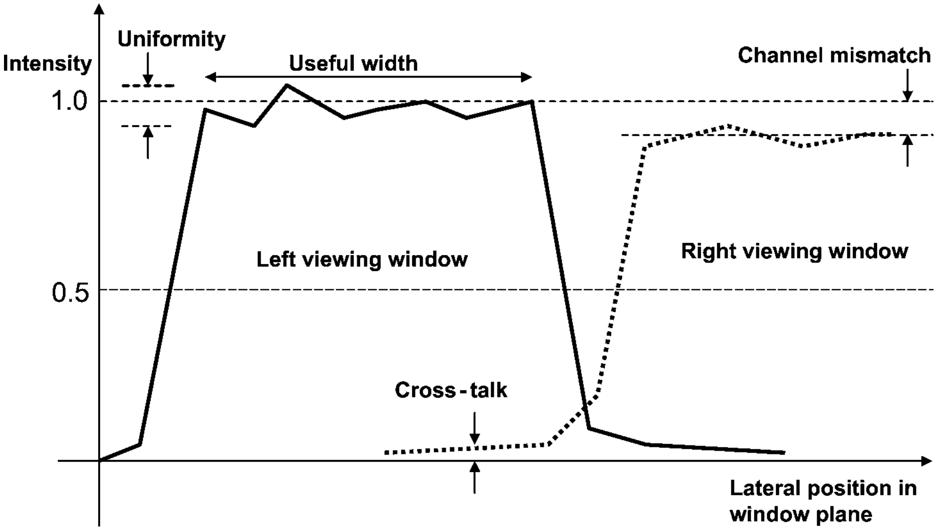


Figure C2.6.11. The characteristics of a viewing window [63].

A systems benefit of wider viewing windows is that it helps relax the tolerances required for window steering and tracking mechanisms in head-tracked displays such as [12, 61]. This is because a wider viewing window allows more time and/or distance before the steering and tracking mechanisms have to respond to user head movement in order to prevent the user moving out of the useful width and seeing a degraded image on the display.

C2.6.3.5 Two-view single-LCD systems

Even with the advantages of a micro-optic design twin-LCD three-dimensional displays have a component cost that must include two LCD elements. This cost is acceptable in some applications when image quality is the key requirement; however, for the mass market, i.e. personal office and home use, it is desirable to find display designs based on a standard single LCD element.

We will group the single-LCD auto-stereoscopic designs by the type of optical element used to generate the viewing windows, beginning with the parallax barrier.

Parallax barrier designs

Typical emissive displays have pixels with diffuse radiance, that is they radiate light equally in all directions. To create a twin-view auto-stereoscopic display, half the pixels must only radiate light in directions seen by the left eye and half the pixels in directions seen by the right eye. The parallax barrier is perhaps the simplest way to do this and works by blocking light using strips of black mask.

The principle of the two-view parallax barrier is illustrated in figure C2.6.12. The left and right images are interlaced in columns on the display and the parallax barrier positioned so that left and right image pixels are blocked from view except in the region of the left and right viewing windows,

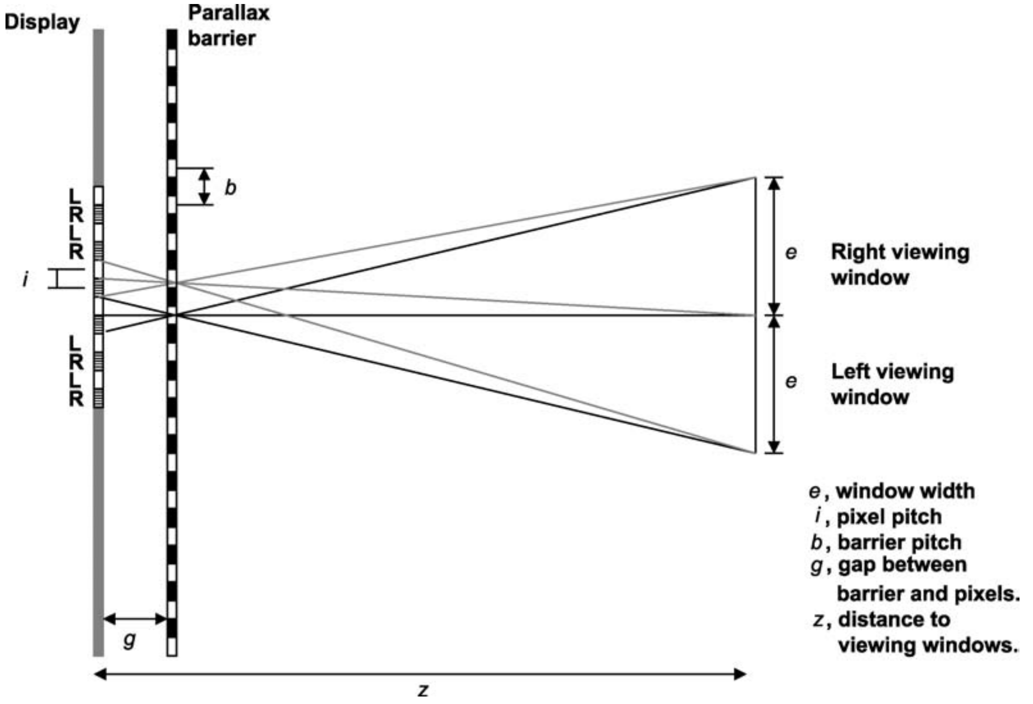


Figure C2.6.12. The principle of the front parallax barrier.

respectively. Although not illustrated the viewing windows repeat in side lobes to each side of the central viewing position and can be used by more than one observer if the optical quality remains high enough.

The pixels and barrier are arranged so the centre of each pair of left and right view pixels is visible at the centre of the viewing windows. The geometry defining the design of the parallax barrier pitch, b, can then be determined from considering similar triangles in figure C2.6.12.

$$\frac{b}{z-g}=\frac{2i}{z} \tag{C2.6.12}$$

which can be rearranged to give:

$$b=2i\left(\frac{z-g}{z}\right). \tag{C2.6.13}$$

The result, equation (C2.6.13), is that the barrier pitch for a two-viewing window display is just less than twice the pixel pitch on the display. This small difference between the pixels and the barrier pitch accounts for the variation in viewing angle between the eyes and the pixels across the display and is often referred to as viewpoint correction.

Viewing distance, z, for the best quality viewing windows is another design factor and again from similar triangles in figure C2.6.12, we can deduce a geometric relationship for this.

$$\frac{i}{g}=\frac{e}{z-g} \tag{C2.6.14}$$

which can be rearranged to give:

$$z=g\left(\frac{e+i}{i}\right). \tag{C2.6.15}$$

The window width is typically set to the average eye separation, $e = 65$ mm, the pixel pitch, i, is defined by the display and the gap, g, between display and barrier is defined by the thickness of the front substrate on the LCD. For example, pixel width might be of the order $i = 0.1$ mm and the gap, including front substrate and polarizer, $g = 1.15$ mm. The result is relatively little control of the closest possible viewing distance and given current LCD substrate thickness many current parallax-barrier-based displays have optimal viewing distances of $z = 750$ mm.

More recent two-dimensional displays could use a substrate such as Corning *Eagle*^{2000} with thickness from 0.4 to 0.63 mm and given a polarizer of thickness 0.2 mm may then be able to reduce viewing distance for a front parallax barrier to $z = 390$ mm. This compares favourably with the typical viewing distance of two-dimensional displays of 300–350 mm although care would be needed to avoid artefacts at the edges of the screen plane where the viewing angle increases with decreasing viewing distance.

Variations on the basic twin-view parallax barrier design and further practical issues are described by Kaplan [29] including a discussion of multi-view parallax barrier displays and aperture design.

Okoshi [40] notes that problems with parallax barriers include the reduced brightness due to blocking the light from pixels, reflection from the glass surface of the parallax barrier and the design of the parallax apertures to avoid diffraction problems. However, these disadvantages have been addressed and recent LCD-based designs overcome the first two problems by using bright light sources and anti-reflection-coated optics. The result is parallax barriers are now widely used for two-view displays such as described [62, 63] and illustrated in figure C2.6.13.

The diffraction problem is more serious but has also recently been addressed. An ideal display would have viewing windows described by a top hat function; however, in practice they have the characteristics

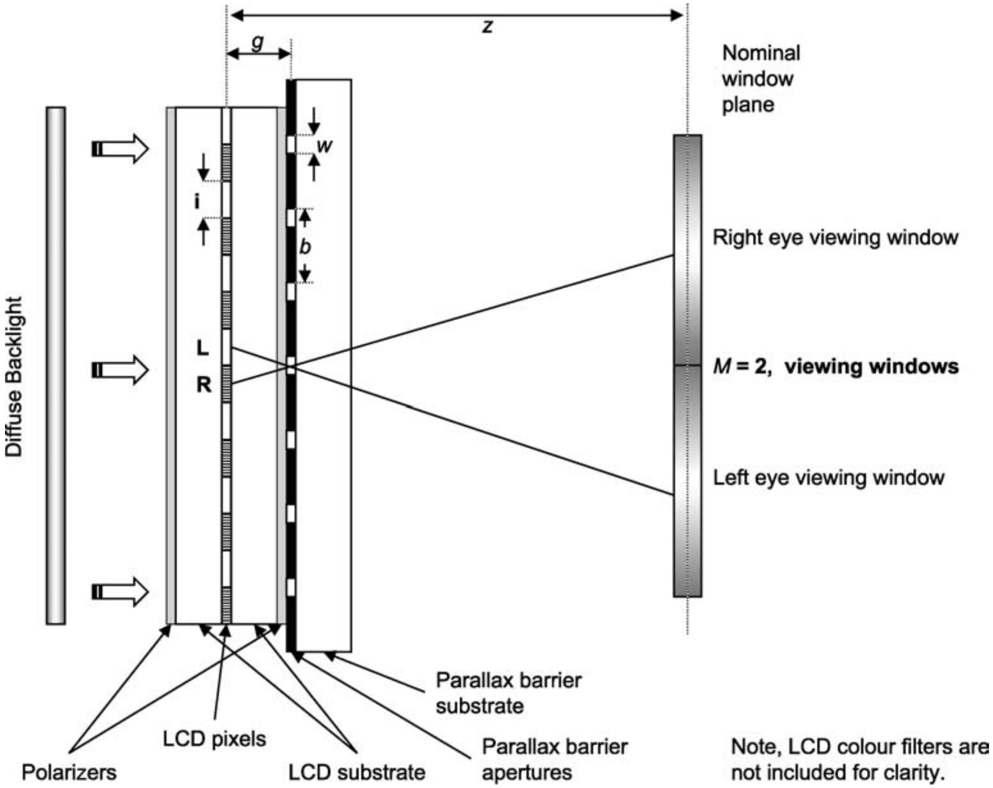

z
g
Nominal
window
plane
w
i
b
Diffuse Backlight
L
R
Right eye viewing window
$M = 2$, viewing windows
Left eye viewing window
Parallax barrier
substrate
LCD pixels
Parallax barrier
apertures
Note, LCD colour filters are
not included for clarity.
Polarizers
LCD substrate

Figure C2.6.13. Detail of a single-LCD front parallax barrier design [62].

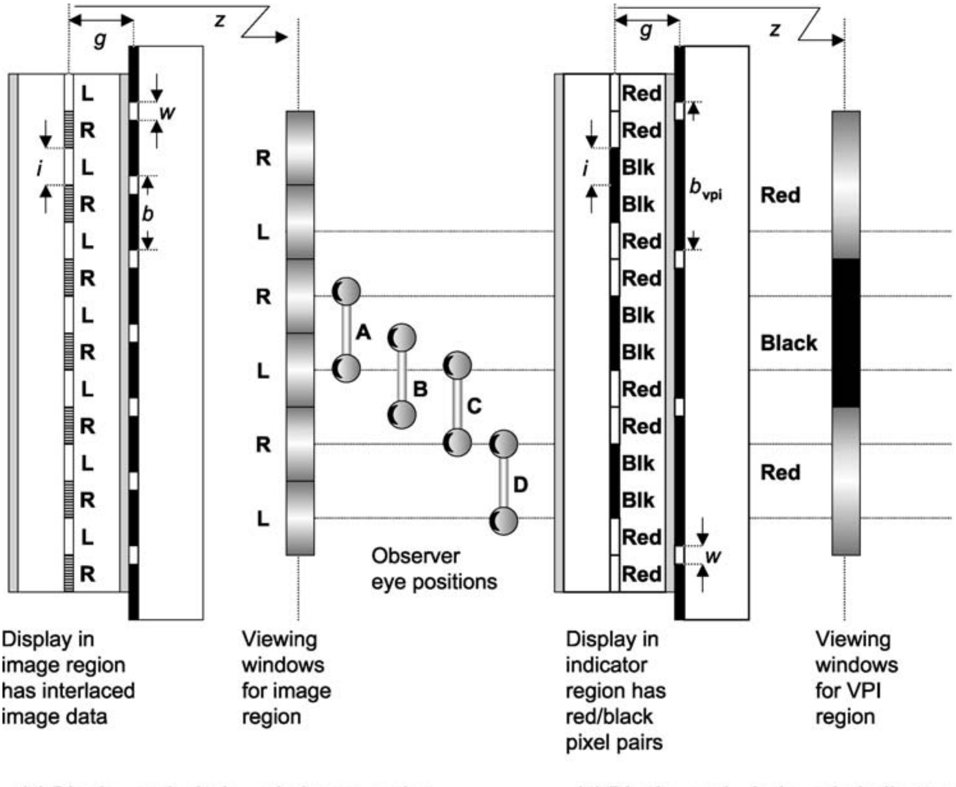
shown in figure C2.6.11. A number of factors determine this and an important one is the detailed design of the parallax barrier apertures, *w*, shown in figure C2.6.13. A wider aperture results in a brighter image but reduces the geometric performance of the aperture and creates less well-defined windows. A narrow aperture results in a less bright image with better window definition; however, too narrow an aperture suffers from diffraction effects which in turn results in less well-defined windows. In both cases, the crosstalk performance, useful width and uniformity of intensity at the viewing window are affected.

A detailed study of the barrier position, aperture design and related diffractive effects is presented in [35, 62]. In [62] a comparison was made between placing the parallax barrier behind and in front of the LCD element. The analysis uses a model of the parallax barrier accounting for Fresnel diffraction and compares this to a set of experimental measurements. Placing the parallax barrier behind the display results in lower crosstalk while placing it in front of the display has very much better intensity uniformity and useful width at the window plane. These factors are decisive for tracking displays and hence the front position was adopted to build a single-LCD observer tracking display [62].

In [35] several apodization modifications to the parallax barrier are analysed; these include soft aperture edges, multiple sub-apertures at aperture edges and combinations of the two techniques. The analysis concluded that choosing the correct apodization can make a substantial improvement to the window profile improving both the crosstalk performance and viewing freedom of the display. In particular, crosstalk of less than 1% is theoretically achievable using an improved parallax element; this is a significant improvement over the value of 3.5% achieved using unmodified apertures. These new studies show it is now possible to overcome the limitations of parallax barriers identified by Okoshi.

A practical problem encountered by users of two-view parallax barrier displays without head tracking is how to find the best viewing position. One reason is the parallax barrier produces not just the central two viewing windows but also repeated lobes to each side of these as illustrated in figure C2.6.14(*a*). An observer in position A will see an orthoscopic image (left image to left eye and right image to right eye) as will an observer in position D. However, an observer in the intermediate position C sees a pseudoscopic image (the left image in the right eye and the right image in the left eye). This causes problems as typically pseudoscopic images show false depth effect and it can be hard for novice observers to determine if they are seeing a correct three-dimensional image or not. A number of devices have been proposed to help observers determine when they are in the correct viewing position; the VPI (viewing position indicator) display described in [62, 63] achieves this by integrating an indicator into the parallax element.

The parallax barrier in the VPI display is divided into two regions: the image region, which is most of the display, and the indicator region, which may cover just the bottom few rows of pixels on the display. The result is shown in figure C2.6.14(*a*) and (*b*), respectively. In the image region the conventional barrier design allows the left and right views to be seen at the nominal viewing position A. In the indicator region, the display shows a pattern of red and black stripes and the barrier design is modified so that the indicator region shows black to both eyes only when the observer is in a position to see an orthoscopic three-dimensional image as at A. If the observer is approaching, as at B, or in, as at C, a pseudoscopic region he will see red in one eye in the indicator region indicating he should move


g
z
L
R
L
R
L
R
L
R
L
R
L
R
L
R
i
w
b
R
L
R
L
R
L
A
B
C
D
Observer eye positions
Display in image region has interlaced image data
Viewing windows for image region
g
z
Red
Red
Blk
Blk
Red
Red
Blk
Blk
Red
Red
Blk
Blk
Red
Red
i
b_{vpi}
w
Red
Black
Red
Display in indicator region has red/black pixel pairs
Viewing windows for VPI region

(*a*) Display and windows in image region. **(*b*) Display and windows in indicator region.**

Figure C2.6.14. The VPI display operation (*a*) in the image region and (*b*) in the indicator region [62].

laterally until returning to the orthoscopic zone. A drawback of the VPI design is that when the observer is in viewing zone D, she can see an orthoscopic image but the indicator will still show red. However, the observer is guaranteed that whenever a black indicator region is seen, he will see an orthoscopic image on the display and this seems a reasonable trade-off.

The indicator region is implemented by using a barrier pitch in the indicator region double that used in the image region. As a result, the VPI display requires little additional design or manufacturing cost and uses only a few lines of pixels to display the appropriate indicator pattern. It has the benefit that once the parallax barrier is aligned for image viewing, the indicator mechanism is automatically aligned. The VPI also works to help guide observers find the best longitudinal viewing position if the aperture width, w, is kept the same in both the image and indicator regions of the parallax element.

A range of designs using parallax barrier optics in combination with LCD elements has been proposed, prototyped and commercialized.

A large range of display designs was developed using parallax barriers [19]. One example uses both a rear and a front parallax barrier with the aim of reducing crosstalk, although no window profile measurements are given to say how successful this was. Because the combination of two parallax barriers reduces display brightness, the rear barrier was mirror coated on the side facing the illuminator to recirculate light. A further design using just a rear parallax barrier places an electronically switchable diffuser between the parallax barrier and the LCD element. This allows instantaneous switching between two-dimensional and three-dimensional modes and if the diffuser is programmable also allows three-dimensional windows to appear on a two-dimensional display. Several designs also combine a window steering mechanism and head tracker to increase lateral viewing freedom; one of these [26] uses an electronically programmable LC parallax barrier.

A design also based on an electronically programmable parallax barrier is described by Perlin [44–46]. A key goal for the design is to steer the viewing windows to track the viewer in three dimensions by varying the pitch and aperture of the parallax barrier in real time. The aim is to generate real time viewpoint correction so the viewer can vary position and still see a three-dimensional image across the whole display surface. The potential benefit of the design is in extending longitudinal movement with respect to the display and it is also capable of accounting for head rotation, which effectively varies the observer's eye separation. The design is relatively complex and before choosing this approach, it would be wise to make a comparison with the longitudinal freedom already available from a fixed aperture display with good quality viewing windows. In practice, realizing the display presents a number of challenges including the optical quality achievable from the programmable parallax element and the speed and latency targets with which the tracking and steering mechanisms need to work.

Lenticular element designs

Lenticular elements used in three-dimensional displays are typically cylindrical lenses arranged vertically with respect to a two-dimensional display such as an LCD. The cylindrical lenses direct the diffuse light from a pixel so it can only be seen in a limited angle in front of the display. This then allows different pixels to be directed to either the left or right viewing windows.

The principle for a two-view lenticular element stereoscopic display is illustrated in figure C2.6.15 and described in [50]. This shows the geometry for a viewpoint-corrected display where the pitch of the lenticular is slightly less than the pitch of the pixel pairs. As with parallax barrier displays the effect of viewpoint correction is to ensure pixels at the edge of the display are seen correctly in the left and right viewing windows. The lenticular pitch needs to be set so that the centre of each pair of pixels is projected to the centre of the viewing windows and this can be found by considering similar triangles where:

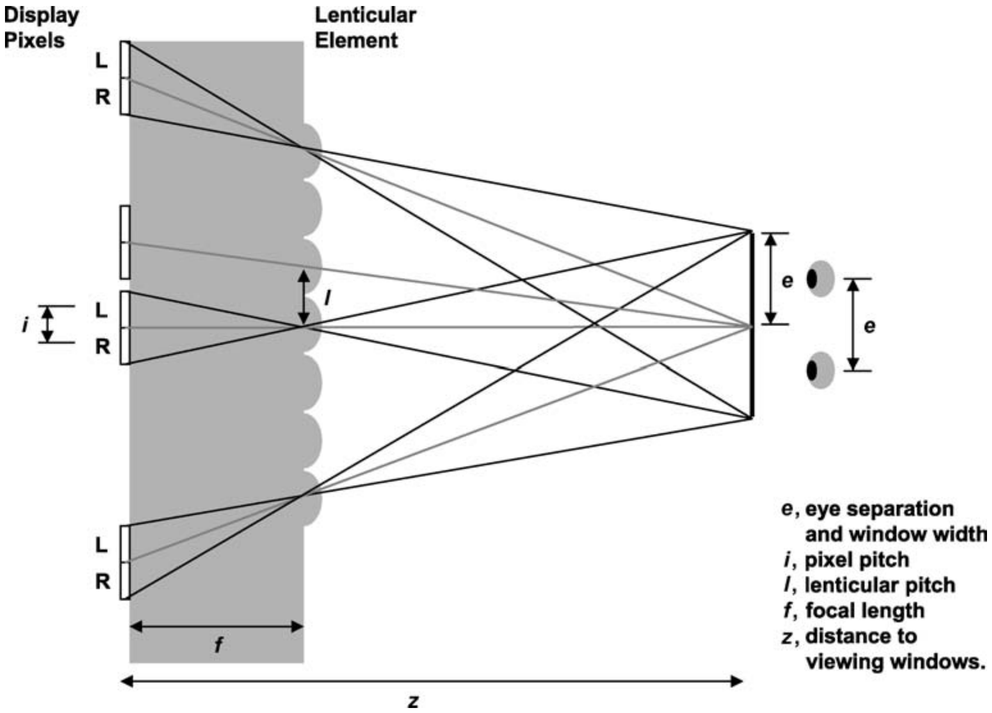

Display Pixels
Lenticular Element
L
R
L
R
L
R
i
l
e
e
f
z
e, eye separation and window width
i, pixel pitch
l, lenticular pitch
f, focal length
z, distance to viewing windows.

Figure C2.6.15. Front lenticular auto-stereoscopic display principle [50].

$$\frac{2i}{z} = \frac{l}{z - f} \tag{C2.6.16}$$

$$l = 2i\left(\frac{z - f}{z}\right). \tag{C2.6.17}$$

Typically, the pixel pitch i is set by the choice of two-dimensional display and the minimum focal length, f, determined in large part by the substrate thickness on the front of the display.

The viewing distance can again be derived from similar triangles:

$$\frac{i}{f} = \frac{e}{z - f} \tag{C2.6.18}$$

which can be easily rearranged to give

$$z = f\left(\frac{e}{i} + 1\right). \tag{C2.6.19}$$

Typically, the window width for a two-view system is taken to be the average eye separation, $e = 65\,\text{mm}$, to give some freedom of movement (up to $e/2$) around the nominal viewing position. Combining this factor with the display-related values of i and f, it may be that there is again little choice over the closest possible viewing distance.

Lenticular elements have been used less often than parallax barriers in recent two-view display designs; one exception is the range of displays designed by the DTI corporation.

The DTI display design described by Eichenlaub [10, 11] uses light guide and lenticular elements behind an LCD display to generate light lines that are functionally equivalent to having a rear parallax barrier. The principle of creating viewing windows using the light lines is shown in figure C2.6.16. The pitch required for the light lines can be calculated using similar triangles as for the parallax barrier example discussed earlier.

$$\frac{b}{z+g} = \frac{2i}{z} \qquad \text{(C2.6.20)}$$

which can be rearranged to give:

$$b = 2i\left(\frac{z+g}{z}\right). \qquad \text{(C2.6.21)}$$

In this case, the pitch of the light lines, b, is slightly larger than twice the pixel pitch to achieve viewpoint correction. Again the gap, g, will determine viewing distance and is likely to be constrained by the substrate glass thickness when using an LCD.

The backlight construction that creates the light lines is shown in figure C2.6.17. A modified light guide uses a series of grooves to generate an initial set of light lines, which are then re-imaged by the lenticular element to form a larger number of evenly spaced light lines in front of the light guide.

A two-dimensional/three-dimensional switching diffuser in front of the lenticular element is made of polymer dispersed liquid crystal (PDLC) which when on is transparent allowing the display to operate in three-dimensional mode. When the PDLC is off it becomes a diffuser, scattering light and preventing the initial set of light lines reaching the lenticular lens. The result is a diffuse illumination for the display, which will operate with similar performance to a normal two-dimensional display. Various size displays

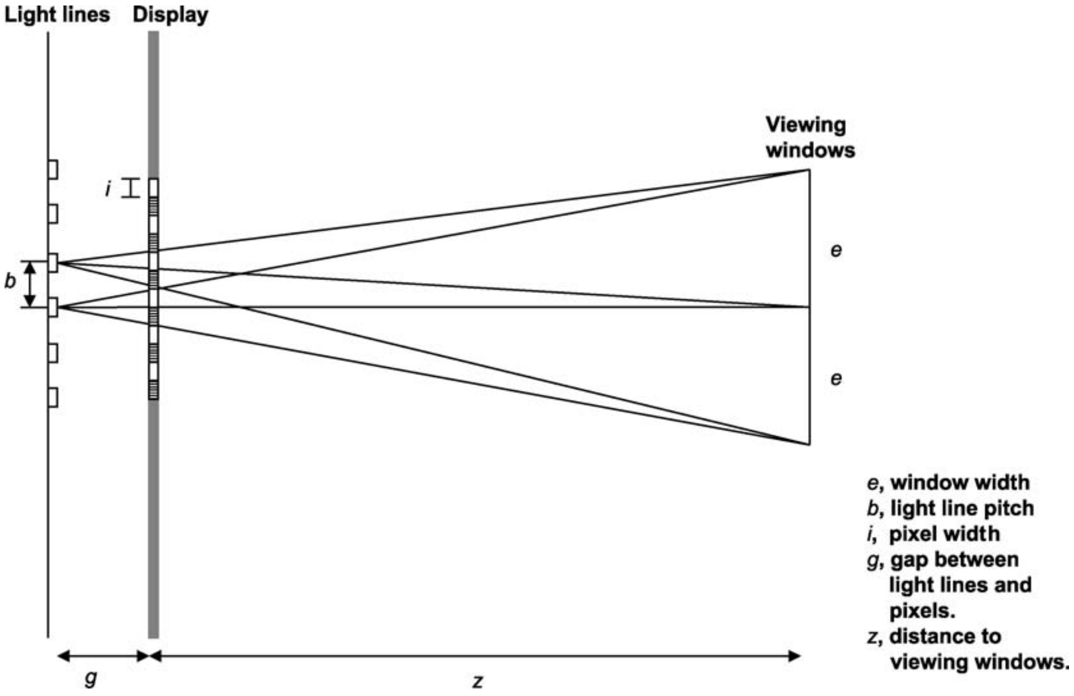


Figure C2.6.16. The geometry of rear parallax illumination by light lines.

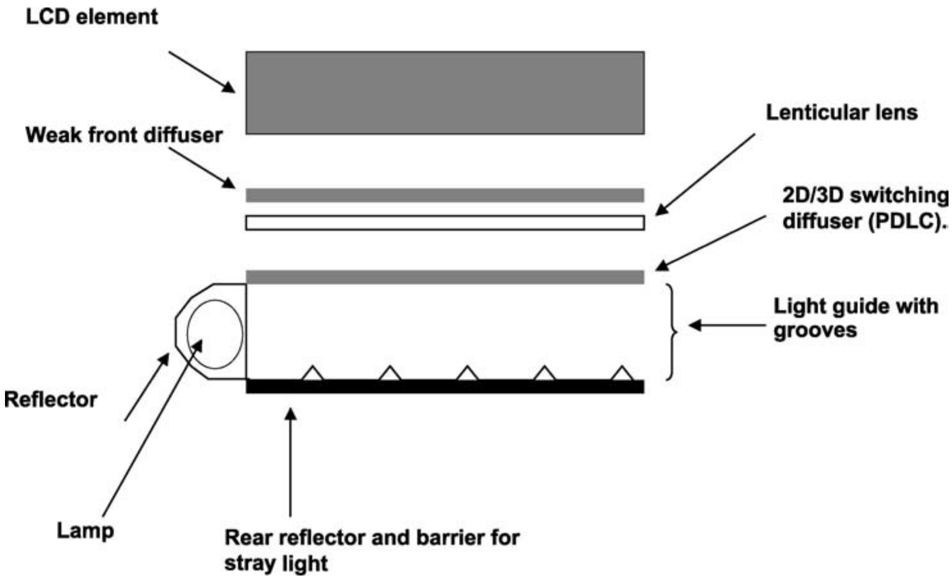


Figure C2.6.17. The DTI compact backlight allowing two-dimensional/three-dimensional illumination.

have been constructed with 5.6 and 12.1 in displays having crosstalk of 3 and 6% and uniformity of 20 and 24%, respectively.

The DTI design has the advantage of being able to electronically switch between two-dimensional and three-dimensional illumination modes as well as being small enough to be used in portable display devices. In addition, there are no optical elements in front of the display surface allowing the observer to directly view the LCD display. Against this are some trade-offs and the three-dimensional mode has higher crosstalk than a well-designed parallax barrier system.

Other designs for single-LCD three-dimensional displays using lenticular optics include [61], [43] and [37, 38].

A novel design using micro-prism elements was proposed [48, 49]. The D4D display uses an array of vertically oriented micro-prisms as the parallax element and the left and right images, vertically interlaced in columns, are directed to two viewing windows by the micro-prisms. A commercial display based on this principle included a head tracking device and both electronic image shifting and mechanical movement of the micro-prisms were investigated as ways to steer the viewing windows.

Micro-polarizer designs

Displays using polarization to create light steering optical elements have been proposed by several groups. The stereoscopic display design described by Faris [14, 15] can also be configured to have an auto-stereoscopic mode by using a series of stacked micro-polarizer elements to create a switchable parallax barrier. However, despite this potential for auto-stereoscopic operation most of the commercial products from VREX have been stereoscopic systems.

Harrold *et al* describe display designs using micro-polarizers in [20, 21]. The design exploits the polarized light output from an LCD element over which is created a patterned retarder array. A final polarizing layer is placed over the retarder array effectively creating a front parallax barrier and hence an auto-stereoscopic display. If the final polarizing layer is constructed so that it is removable, the display can be mechanically switched between a two-dimensional display mode and an auto-stereoscopic three-dimensional display mode (figure C2.6.18).

Key to the success of this design is the construction of the patterned retarder array to an accuracy of better than 1 part in 2000 for the 13.8 in XGA display prototype. This was achieved using a process based on standard LCD manufacturing techniques to create a manufacturable patterned retarder array that is front mounted onto the LCD element.

A stereoscopic display design is also described by Harrold in [20] where the patterned retarder and polarizer are constructed inside the LCD element to avoid the parallax problems of the Faris design [14, 15]. A prototype LC cell demonstrated the feasibility of this approach.

A micro-polarizer display described by Benton [1, 2] uses a combination of polarization and bulk optics to create two viewing windows that can be steered electronically if a suitable head tracker is available. An LCD panel with the analysing polarizer removed acts as an electronically programmable polarizing light source: light coming from the light source LCD will be either rotated at 90° or not rotated. An illumination pattern of two blocks of light is displayed on the light source LCD, each polarized differently. A micropolarizer array arranged as rows behind an image LCD display allows alternate rows of image to be illuminated by differently polarized light and hence appear in the viewing windows for the left and right eyes. A large lens after the LCD produces an image of the viewer-tracking LCD (polarized light source) at the intended viewing distance of about 1 m creating the two viewing windows.

Benton notes there can be problems with the lens (a Fresnel lens) creating Moire patterns in association with the image display LCD. In common with many auto-stereoscopic displays the viewer has to be at or close to the nominal viewing distance, which at 1 m is significantly further than typical

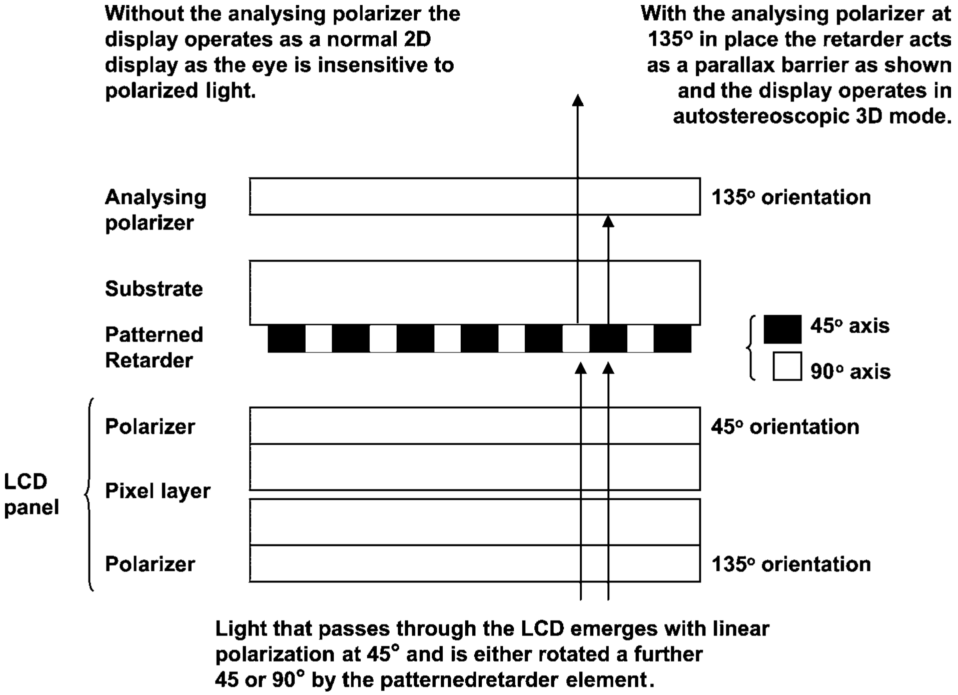

Without the analysing polarizer the display operates as a normal 2D display as the eye is insensitive to polarized light.
With the analysing polarizer at 135° in place the retarder acts as a parallax barrier as shown and the display operates in autostereoscopic 3D mode.
Analysing polarizer
135° orientation
Substrate
Patterned Retarder
45° axis
90° axis
LCD panel
Polarizer
45° orientation
Pixel layer
Polarizer
135° orientation
Light that passes through the LCD emerges with linear polarization at 45° and is either rotated a further 45 or 90° by the patternedretarder element.

Figure C2.6.18. The Sharp micro-polarizer display with two-dimensional/three-dimensional switching capability [20].

two-dimensional display viewing distances. No measurements of crosstalk or window brightness uniformity are given.

Holographic elements

Holographic optical elements (HOEs) have been used [51, 52] to create three-dimensional displays in conjunction with LCD elements. When illuminated the HOE acts to form the viewing windows. The HOE is arranged in horizontal strips to reconstruct a real image of a diffuse illuminator; the strips are arranged so alternate strips reconstruct left and right viewing windows. When placed behind a display with two horizontally interlaced images, the observer will see an auto-stereoscopic image.

A number of practical problems in the optical design are discussed in [51] and in particular colour fringing due to the diffractive nature of the HOE could prove difficult to overcome. Otherwise, this design has several advantages and can be modified to track users by moving the light source and also constructed so that it can be switched between two-dimensional and three-dimensional using a modified light source.

C2.6.3.6 Multi-view systems

The viewing freedom of a three-dimensional display is a key requirement in certain applications, for example public information kiosks, where ease of viewing is needed to attract and retain the attention of passers-by. Multi-view systems, as in figure C2.6.19, provide viewing freedom by generating multiple simultaneous viewing windows of which an observer sees just two at any time. Multi-view systems can also support more than one observer if enough horizontal viewing freedom is available.

Bulk optic multi-view displays have been developed and are reviewed in the literature [7, 36]. The display was designed to use temporal multiplexing of the view images and because the basic switching speed and interface bandwidth of LCD displays were not sufficient, this led to the use of high-speed CRT technology.

Micro-optic multi-view designs using standard two-dimensional displays have been proposed where the images are spatially multiplexed. The Heinrich-Hertz-Institut has a well-established programme investigating lenticular three-dimensional displays and Borner [4] describes a number of multi-view designs.

The principle for a multi-view LCD display using a front lenticular element, similar to the two-view lenticular design described previously, is illustrated in figure C2.6.20. This shows a five-view lenticular

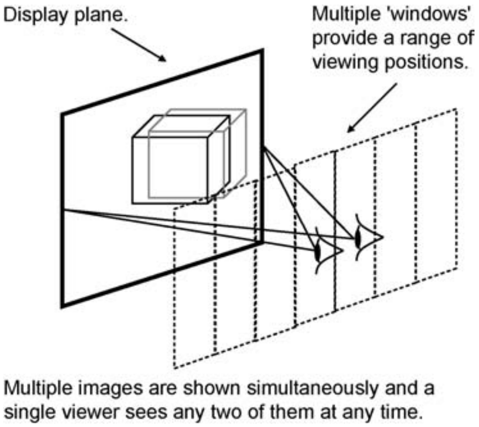

Display plane.
Multiple 'windows' provide a range of viewing positions.
Multiple images are shown simultaneously and a single viewer sees any two of them at any time.

Figure C2.6.19. Multi-view displays create multiple viewing windows.

display, where each pixel in every group of five pixels is directed to a different viewing window. As with the two-view displays the system should be viewpoint corrected so that the viewing windows are aligned with pixels across the whole display.

To use the display, five images are sliced vertically into columns and interlaced appropriately. The images will then be visible separately in the five viewing windows V1–V5 in figure C2.6.20. The viewing windows can be designed as shown so pairs of image separated by one image, for example V2 and V4, are seen simultaneously by the left and right eyes and if these form a stereo pair, then an observer sees an image with stereoscopic depth. In addition, if the observer moves laterally they can see a different pair of images, for example V3 and V5, and therefore a different stereoscopic view of the scene.

Using a similar geometrical argument as for two-view lenticular displays, the pitch of the lenses can be determined by:

$$l = N_{\mathrm{v}} i \left(\frac{z - f}{z} \right) \tag{C2.6.22}$$

where N_{v} is the number of viewing windows required.

There are several drawbacks to the basic multi-view approach that are particularly apparent when electronic displays are used [55]. The first is there is a black mask between LCD pixels and this is imaged into dark lines between each view window which is distracting to observers when the eye crosses a window boundary. Also images with any significant depth will result in an image-flipping artefact as the observer moves the eye across one view window and into the next. Finally as more views are used the horizontal resolution of the images decreases rapidly. To overcome these problems, a new approach to multi-view LCD display was proposed [55].

Several multi-view systems based on lenticular micro-optics and single LCD displays were proposed [54–56]. A significant step forward was made by positioning the lenticular array at an angle to the LCD

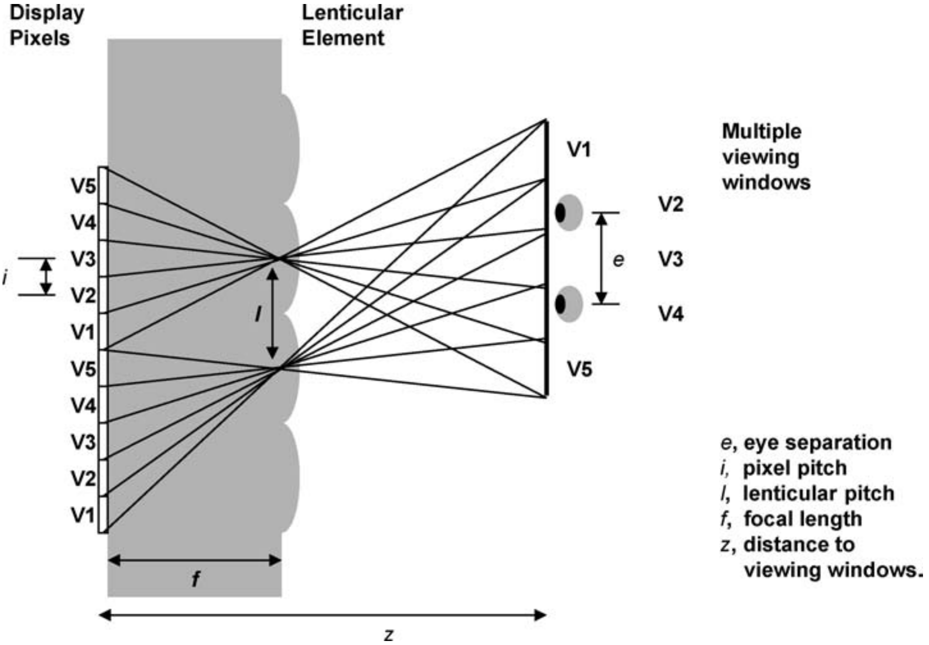


Figure C2.6.20. The principle of a multi-view front lenticular auto-stereoscopic display.

pixel array; this mixed adjacent views reducing image flipping problems and spreading the effect of the black mask making it less visible. The other benefit of this design is that each view has a better aspect ratio; rather than splitting the display horizontally into many views both horizontal and vertical directions are split.

The arrangement of one lenticule and the underlying pixels in the slanted lenticular design is shown in figure C2.6.21. The slanted lenticular arrangement means that all pixels along ‘a’ line such as a will be imaged in the same direction. In this case all view three pixels are seen in the same direction. The arrangement shown allows seven views to be interlaced on the display and imaged in different directions by the lenticule. As the eye moves from position ‘a’ to ‘b’ the eye sees a gradual transition from view 3 to view 4. At most viewing positions the eyes will see a combination of more than one view; while this inherent crosstalk limits the depth that can be shown on the display it does hide the transition between views at boundaries and blurs the appearance of the black mask so that it is no longer an obvious visual artefact. For the seven-view display described in [56] the magnification of the lenticules is designed so that a viewer at a distance of approximately 700 mm from the display sees views 3 and 5 in left and right eyes respectively, i.e. views separated by one view form a stereo pair.

An alternative design where the pixels are slanted instead of the lenticular element is described in [57]. However, such a major change to LC display design is unlikely to happen unless there is a substantial worldwide market for three-dimensional displays or an advantage of slanted pixels for two-dimensional LCD operation is found.

The multi-view display design [33] adopts a similar solution, citing an earlier reference [59] as the source of the idea for using a lenticular slanted with respect to the vertical image axis. This display generates nine viewing regions, through which the user can see nine equal resolution images. Based on an SXGA (1280 × 1024) LCD display this results in each viewing window image having a two-dimensional resolution of 426 by 341 pixels.

Experience with lenticular optics [61] suggests displays based on lenticular optics have to make additional design trade-offs. An important one is the difficulty of anti-reflection coating the lenses, which can lead to distracting reflections on the display surface. Another is that scattering of light in the lenses generates a visible artefact looking to the user like a light grey mist present throughout the three-dimensional scene.

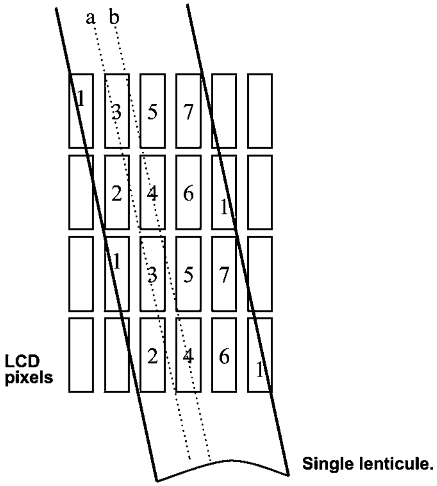


Figure C2.6.21. The slanted arrangement of the lenticular lens and pixels in the multi-view display [55].

To summarize multi-view displays:

- Temporally multiplexed displays with high resolution per view suffer a number of drawbacks: they need high-speed display elements and high-bandwidth image generation and interface circuits. This seems likely to delay their widespread adoption in personal three-dimensional display applications.
- Spatially multiplexed designs have lower resolution per view than twin-view displays and recent designs build in crosstalk limiting the three-dimensional depth. Despite this they are attractive commercially because of the benefit of viewing freedom they provide and their relatively low cost and manufacturability.

A solution for the future is to build a system with an intermediate number of views, say three, not requiring mechanical view steering and use a head tracking device to keep the images up to date with the observer's head position. One such system, known as PixCon, is described in [61] and another design is presented in [48]. A similar idea for using view switching in a twin-view system was proposed in [50]. A prerequisite for this is low-cost, accurate, observer head tracking and some good progress is being made in this area [25].

C2.6.4 Three-dimensional display performance and use

C2.6.4.1 Comparing perceived depth reproduction

Perceived depth reproduction is the single most important reason for building three-dimensional displays but system characteristics in this respect are rarely reported in the literature. In this section, we consider three generic designs, the twin-LCD and single-LCD two-view systems and a single-LCD multi-view system, and analyse their ability to reproduce perceived depth. Similar real examples of these designs are the twin-LCD display [61], the single-LCD VPI display [62] and the nine-view multi-view display, but our discussion abstracts from the details of specific display implementations for clarity. We compare the ability of the three generic designs to reproduce depth to each other and to the performance of the human eye; we also consider the demands on the graphics and imaging systems supplying the displays with content.

The generic three-dimensional display designs are assumed to be based on the same underlying LCD element, a 1280 by 1024 pixel display with a horizontal pixel width of $i = 0.3\,\text{mm}$ approximating an 18.8 in diagonal SXGA display. The three-dimensional displays can then be characterized by the effective pixel width in the image seen by one eye:

- The twin-LCD twin-view display has two overlaid images and the pixel width in each view is the same as the base panel at $i = 0.3\,\text{mm}$.
- The single-LCD twin-view display has two horizontally interlaced images and the pixel width in each view is double the base panel at $i = 0.6\,\text{mm}$.
- The single-LCD multi-view display has nine views, interlaced horizontally and vertically, and the pixel width in each view is triple the base panel at $i = 0.9\,\text{mm}$.

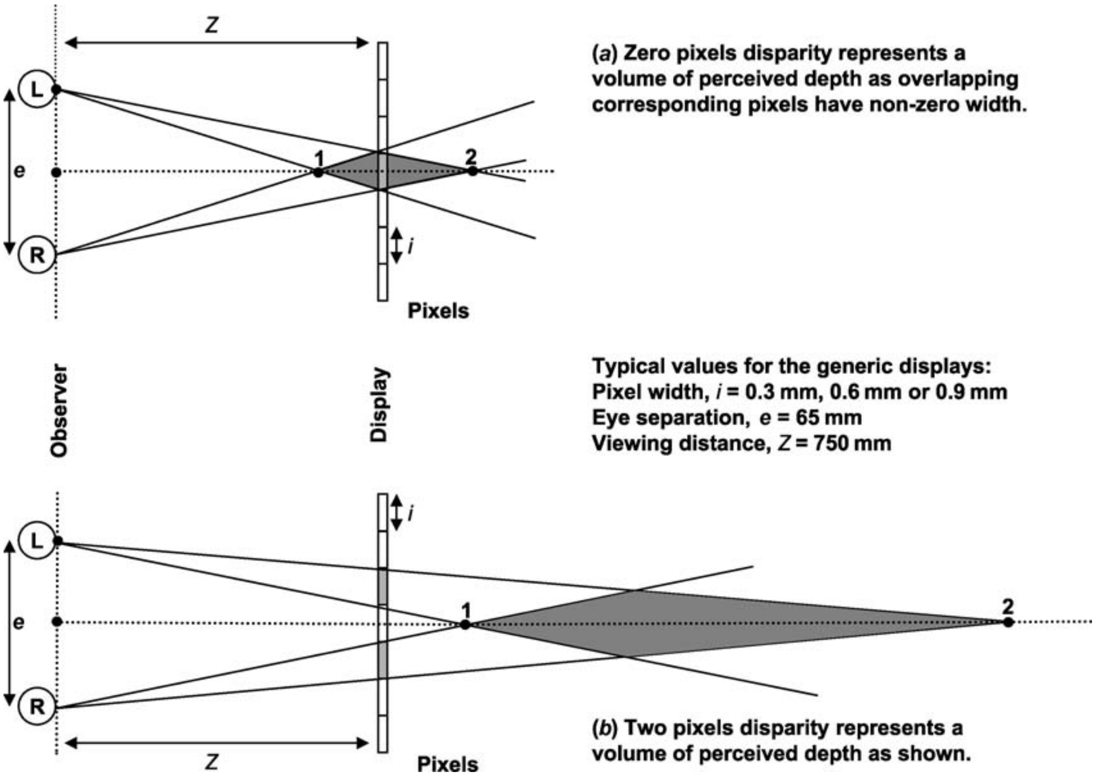
We assume the latter two displays overlay the left and right eye images to simplify discussion, but note in practice it will be necessary to consider the exact interlacing of RGB components.

The following set of characteristics provides a basis for comparing display designs. Our aim is to capture the characteristics that are important in the human perception of three-dimensional displays.

Total display resolution: However a stereoscopic display is designed to provide views to each eye, the total display resolution, i.e. the sum of all pixels in all views, largely determines the computational effort required to generate the images for display and the bandwidth required in interface circuits. Displays which require image interlacing will also require additional functionality in interface circuits as pixels from different views typically need to be interlaced at the RGB component level. Bandwidth requirements can be determined from total display resolution and the desired frame rate.

Resolution per view: The resolution per view is a key characteristic of a three-dimensional display. Having stereo three dimensions does not replace the need for high spatial resolution and anyone used to 1280 × 1024 monoscopic displays will notice the step down when dividing these pixels between two or nine views. However, a three-dimensional display can often look better than a monoscopic display with the same resolution as a single view on the three-dimensional display because the brain integrates the information received from the two views into a single image.

Perceived depth voxels: As shown in figure C2.6.22, a pair of corresponding pixels in the left and right images represents a volume of perceived depth; we will call this a stereoscopic voxel or voxel as in [24]. Of particular interest is the depth of a voxel that a display can represent for a given screen disparity between corresponding pixels. We can use this to compare the depth representation abilities of different displays in depth and to compare displays with the ability of the eye to perceive depth. The perceived voxels are arranged in planes from in front to behind the display; as they recede from the viewer the cells increase in depth [8, 24].



(*a*) Zero pixels disparity represents a volume of perceived depth as overlapping corresponding pixels have non-zero width.

Typical values for the generic displays:
Pixel width, *i* = 0.3 mm, 0.6 mm or 0.9 mm
Eye separation, *e* = 65 mm
Viewing distance, *Z* = 750 mm

(*b*) Two pixels disparity represents a volume of perceived depth as shown.

Figure C2.6.22. The perceived depth represented by corresponding pixels of 0 and 2 pixels screen disparity.

The depth span of a voxel can be found by using equations (C2.6.10) and (C2.6.11) as appropriate to calculate difference in depth of points 1 and 2 in figure C2.6.22.

Consider zero pixels disparity as in figure C2.6.22(*a*). For point 1, a pixel width $i = 0.3$ mm implies screen disparity of $d = -0.3$ mm and assuming $z = 750$ mm and $e = 65$ mm then the perceived depth in front of the screen plane is:

$$p = \frac{750}{\left(\frac{65}{|-0.3|}\right) + 1} = 3.45\,\text{mm}. \tag{C2.6.23}$$

For point 2 the screen disparity is $d = 0.3$ mm and the perceived depth behind the screen plane is:

$$p = \frac{750}{\left(\frac{65}{|0.3|}\right) - 1} = 3.48\,\text{mm}. \tag{C2.6.24}$$

Therefore, the total perceived voxel depth in this case is 6.93 mm. This is the perceived depth represented by corresponding pixels with zero disparity at the screen plane. In practice, it tells us this display cannot reproduce a depth difference between objects at the screen plane of less than 6.93 mm. Results of similar calculations for all three generic displays are given in figure C2.6.24.

Perceived depth range: The perceived depth range, that is the nearest and furthest points a display can reproduce, is of interest. Geometrically, this can be calculated from the maximum screen disparity available; however, for most displays of any size the geometric range is much more than can be viewed comfortably by the majority of observers. Instead, it is important to determine the comfortable perceived depth range experimentally and for our discussion we adopt results reported in [27]. This suggests a comfortable working range for the majority of people is from 100 mm in front to 100 mm behind the display surface and this range could probably be extended to 200 mm in front and 500 mm behind and still be comfortable to view for the majority of observers. We take the ±100 mm range for our calculations here without affecting the generality of the discussion.

Stereoscopic resolution: Identifying the comfortable working range of perceived depth on a display also allows us to define the resolution of perceived depth within this range. Perceived depth voxels of equal screen disparity form planes of voxels parallel to the display surface as illustrated in figure C2.6.23. We will define stereoscopic resolution to be the number of planes of voxels within the range of ±100 mm.

Stereoscopic resolution can be calculated identically for each of the generic displays, which have the same viewing distance, by finding the screen disparity, d, that generates voxels at ±100 mm. The sum of these values is then divided by the width of a stereoscopic pixel, i, on the display in question.

The table in figure C2.6.24 shows values of the characteristics discussed here for the three generic displays. Not surprisingly the twin-LCD display with the most pixels per view has the best results for depth reproduction with an ability to reproduce depth differences of 7 mm at the screen plane and a stereoscopic resolution of 60 planes of depth in the working depth range ±100 mm. However, the eye is much better at perceiving depth than the best display is at reproducing it with a minimum detectable depth difference of 0.84 mm and an equivalent stereoscopic resolution of 240 planes of depth in the working range ±100 mm.

This difference suggests significant improvements are still possible to the depth reproduction characteristics of stereoscopic displays. It is also important to keep in mind when using the displays if depth judgement is critical to task performance. In the next section, we briefly review how to create images that account for the available working depth range and resolution of three-dimensional displays.

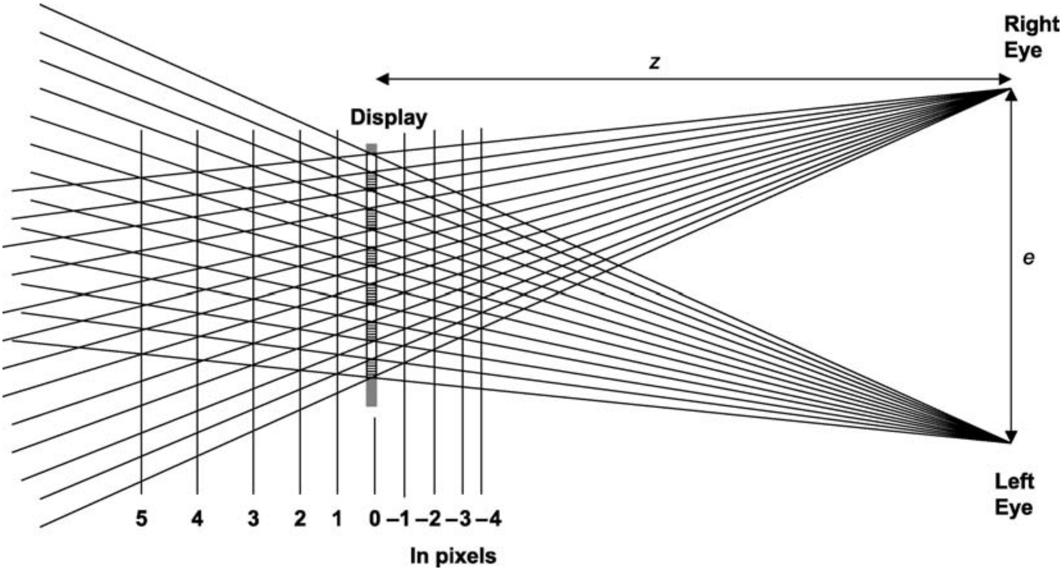
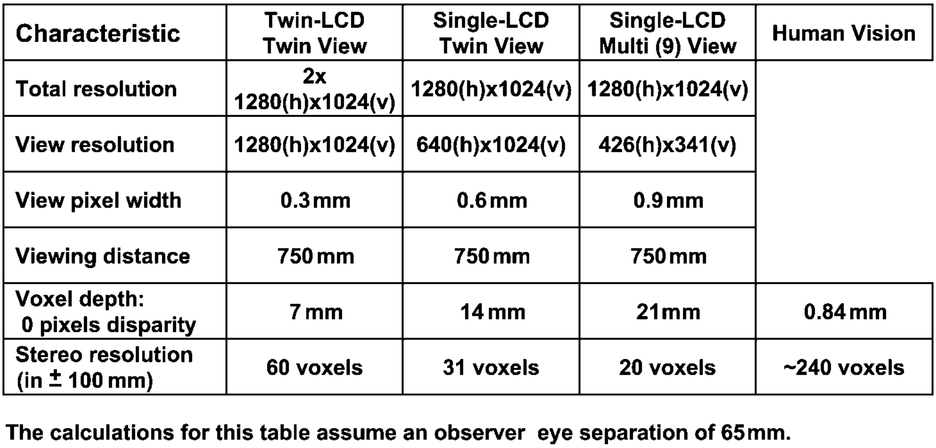


Figure C2.6.23. Stereoscopic resolution is defined by planes of stereoscopic voxels.



Characteristic	Twin-LCD Twin View	Single-LCD Twin View	Single-LCD Multi (9) View	Human Vision
Total resolution	2x 1280(h)x1024(v)	1280(h)x1024(v)	1280(h)x1024(v)	
View resolution	1280(h)x1024(v)	640(h)x1024(v)	426(h)x341(v)	
View pixel width	0.3 mm	0.6 mm	0.9 mm	
Viewing distance	750 mm	750 mm	750 mm	
Voxel depth: 0 pixels disparity	7 mm	14 mm	21 mm	0.84 mm
Stereo resolution (in ± 100 mm)	60 voxels	31 voxels	20 voxels	~240 voxels

The calculations for this table assume an observer eye separation of 65 mm.

Figure C2.6.24. Table comparing characteristics of the generic displays and the eye.

C2.6.4.2 Perceived depth control and image generation

As discussed three-dimensional displays have limits on the comfortable perceived depth range they can reproduce. This results in a working volume of space around the display plane that content producers can use to present a scene in. The working volume available on the display is unlikely to match the volume in the scene being captured. As a result, several approaches to mapping depth from a scene onto the available depth range of the target display have been proposed.

Traditionally, this has involved a discussion of whether to set cameras at eye separation or not and whether camera axes should be parallel or verging. However, recent methods approach the problem as a mapping of a volume of scene space onto the available working volume on the target display.

These methods automatically calculate stereoscopic camera parameters given a camera position, scene volume to capture and target display specifications. Wartell describes one approach in [58] while a simpler and more general method is given by Jones [27]. These have significant benefits in ease of image generation for content creators and guarantee that depth mappings will be geometrically consistent even on head-tracked displays. The result is stereoscopic images should no longer be produced with excessive perceived depth or unwanted distortions.

Despite the long history of stereo image generation, it is only recently that new technology, in the form of three-dimensional computer graphics and digital camera technology, has been able to give enough control over the image creation process to use these methods. The methods are particularly important to apply when creating images for testing and improving a display design because poorly created images, or even just images produced correctly for a different three-dimensional display, can cause the highest-quality display to become uncomfortable to view.

C2.6.5 Summary

Advances in micro-optics, display technologies and computing systems are combining to produce an exciting range of new opportunities for three-dimensional display designers. To achieve a good three-dimensional display design requires a systems approach combining optical, electrical, mechanical and digital imaging skills along with an understanding of the mechanism of binocular vision.

The characteristics and geometry of binocular vision define limits on the maximum range of binocular vision and the minimum depth differences it is possible to perceive in the natural world. Because the perception of depth is relative to the current fixation point, binocular vision is best suited to making relative depth judgements between objects.

Stereoscopic images do not provide the same stimulus to the eyes as the natural world and the implications of this affect three-dimensional display design and use. In particular, while the eyes verge to fixate different depths in a stereo image the eye's accommodation must keep the image plane, rather than the fixation point, in focus. This places measurable limits on how much perceived depth is comfortable to view on a particular three-dimensional display.

As well as the stereoscopic depth cue the brain uses many two-dimensional depth cues to help it understand depth information in a scene. Therefore, the first aim for a three-dimensional display design needs to be to keep the same basic image quality as a two-dimensional display including values of brightness, contrast, spatial resolution and viewing freedom.

We have introduced two-view and multi-view auto-stereoscopic display designs based on micro-optic elements including: parallax barriers, lenticular arrays, micro-polarizers and holographic optical elements. These provide different trade-offs in cost, system complexity and performance. Key characteristics that define the performance of different displays include:

- perceived voxel depth at zero disparity, i.e. minimum reproducible depth at the screen plane;
- stereoscopic resolution, i.e. the number of discrete voxel planes in ±100 mm depth;
- viewing window characteristics particularly inter-channel crosstalk and uniformity.

As three-dimensional display quality continues to improve, it becomes increasingly important to consider the quality of the stereoscopic images used to evaluate displays. This requires the adoption of new methods for image generation based on an improved understanding of the human perception of stereo images to define the mapping of depth from a scene onto the working depth range available on the three-dimensional display.

References

[1] Benton S A, Slowe T E, Kropp A B and Smith S L 1999 Micropolarizer-based multiple-viewer auto-stereoscopic display *Proc. SPIE* **3639**

[2] Benton S A 2002 *Auto-Stereoscopic Display System* US 6 351 280 (filed November 1998)

[3] Blakemore C 1970 The range and scope of binocular depth discrimination in man *Physiology* **211** 599–622

[4] Borner R 1993 Autostereoscopic 3D-imaging by front and rear projection on flat panel displays *Displays* **14**

[5] Cumming B G and DeAngelis G C 2001 The physiology of stereopsis *Annu. Rev. Neurosci.* **24** 203–238

[6] Dodgson N 1997 Autostereo displays: 3D without glasses *Proc. EID* **EID97**

[7] Dodgson N A 2002 Analysis of the viewing zome of multi-view auto-stereoscopic displays *Proc. SPIE* **4660**

[8] Diner D B and Fender D H 1993 *Human Engineering in Stereoscopic Display Devices* (New York: Plenum)

[9] Eichenlaub J 1997 A compact, lightweight, 2D/3D auto-stereoscopic backlight for games, monitor and notebook applications *Proc. SPIE* **3012**

[10] Eichenlaub J 1998 A compact, lightweight, 2D/3D auto-stereoscopic backlight for games, monitor and notebook applications *Proc. SPIE* **3295**

[11] Eichenlaub J and Gruhlke R Reduced thickness backlighter for auto-stereoscopic display and display using the backlighter *US Patent* 5 897 184

[12] Ezra D, Woodgate G, Harrold J, Omar B, Holliman N and Shaprio L 1995 New auto-stereoscopic display system *Proc. SPIE* **2409**

[13] Ezra D, Woodgate G J and Omar B 1998 Autostereoscopic directional display apparatus *US Patent* 5 726 800 (priority from December 17th 1992)

[14] Faris S M 1994 Novel 3D-stereoscopic imaging technology *Proc. SPIE* **2177**

[15] Faris S M 1993 Multi-mode stereoscopic imaging system *US Patent* 5 264 964

[16] Glassner A 1995 *Principles of Digital Image Synthesis* (San Mateo, CA: Morgan Kaufmann)

[17] Gooding L, Miller M E, Moore J and Kim S 1991 The effect of viewing distance and disparity on perceived depth *Proc. SPIE* **1457**

[18] Goldstein E B 2002 *Sensation and Perception* 6th edn (Belmont, CA: Wadsworth)

[19] Hamagishi G, Sakata M, Yamashita A, Mashitani K, Inoue M and Shimizu E 2001 15″ high resolution non-glasses 3-D display with head-tracking system *Trans. IEE, Jpn* **121-C**

[20] Harrold J, Jacobs A M S, Woodgate G J and Ezra D 1999 3D display systems hardware research at Sharp Laboratories of Europe: an update *Sharp Tech. J.*

[21] Harrold J, Jacobs A, Woodgate G J and Ezra D 2000 Performance of a convertible, 2D and 3D parallax barrier auto-stereoscopic display *Proceedings SID, 20th International Display Research Conference, September 2000, Palm Beach, FL*

[22] Hattori T, Ishigaki T, Shimamoto K, Sawaki A, Ishiguchi T and Kobayashi H 1999 An advanced auto-stereoscopic display for G7 pilot project *Proc. SPIE* **3639**

[23] Helmholtz H 2000 *Treatise on Physiological Optics 1867* 1924 edn (reprinted Bristol: Thoemmes)

[24] Hodges L F and Davis E T 1993 Geometric considerations for stereoscopic virtual environments *Presence* **2**

[25] Holliman N, Hong Q, Woodgate G and Ezra D 2000 Image tracking system and method and observer tracking auto-stereoscopic display *US Patent* 6 075 557

[26] Inoue M, Hamagishi G, Sakata M, Yamashita A and Mahitani K 2000 Non-glasses 3-D displays by shift-image splitter technology *Proceedings 3D Image Conference 2000, Tokyo, July 2000*

[27] Jones G, Lee D, Holliman N and Ezra D 2001 Controlling perceived depth in stereoscopic images *Proc. SPIE* **4297A**

[28] Julesz B 1971 *Foundations of Cyclopean Perception* (Chicago, IL: University of Chicago Press)

[29] Kaplan S 1952 Theory of parallax barriers *J. SMPTE* **59**

[30] Lane B 1982 Stereoscopic displays *Proc. SPIE* **0367**

[31] Langlands N 1926 Experiments on binocular vision *Trans. Opt. Soc.* **27** 4–82

[32] Lipton L 1982 *Foundations of Stereoscopic Cinema* now available electronically (Princeton, NJ: Van Nostrand-Reinhold)

[33] Lipton L 2002 Synthagram: auto-stereoscopic display technology *Proc. SPIE* **4660**

[34] McAllister D F 1993 *Stereo Computer Graphics and Other True 3D Technologies* (Princeton, NJ: Princeton University Press)

[35] Montgomery D J, Woodgate G J and Ezra D 2001 Parallax barrier for an auto-stereoscopic display *GB Patent* 2 352 573

[36] Moore J, Dodgson N, Travis A and Lang S 1996 Time-multiplexed color auto-stereoscopic display *Proc. SPIE* **2653**

[37] Morishima H, Nose H and Taniguchi N Stereoscopic image display apparatus *US Patent* 6 160 527

[38] Nose H 1997 Rear-lenticular 3D-LCD without eyeglasses *O plus E* no. 217, pp 105–109

[39] Ogle K N 1964 *Researches in Binocular Vision* (London: Hafner)

[40] Okoshi T 1976 *Three-Dimensional Imaging Techniques* (New York: Academic)

[41] Pastoor S 1991 3D-television: a survey of recent research results on subjective requirements *Signal Process.: Image Commun.* **4**

[42] Pastoor S 1995 *Human Factors of 3D Imaging* Web document distributed by Heinrich-Hertz-Institut, Berlin

[43] Pastoor S and Wopking M 1997 3-D displays: a review of current technologies *Displays* **17**

[44] Perlin K, Paxia S and Kollin J S 2000 An auto-stereoscopic display *Proceedings ACM Siggraph Conference, July 2000*

[45] Perlin K 2001 Displayer and method for displaying *US Patent* 6 239 830 (filed May 1999)

[46] Perlin K, Poultney C, Kollin J S, Kristjansson D T and Paxia S 2001 Recent advances in the NYU auto-stereoscopic display *Proc. SPIE* **4297**

[47] Schiffman H R 2000 *Sensation and Perception: an Integrated Approach* 5th edn (New York: Wiley)

[48] Schwerdtner A and Heidrich H 1998 The Dresden 3D Display (D4D) *Proc. SPIE* **3295**

[49] Schwerdtner A and Heidrich H 1998 Optical system for the two and three dimensional representation of information *US Patent* 5 774 262 (filed Germany 1993)

[50] Susumu I, Nobuji T and Morito I 1990 Technique of stereoscopic image display *EP Patent* 0 354 851 (filed Japan August 1988)

[51] Trayner D and Orr E 1997 Developments in auto-stereoscopic displays using holographic optical elements *Proc. SPIE* **3012**

[52] Trayner D and Orr E Direct View holographic auto-stereoscopic displays *Proceedings of the Fourth UK VR-SIG, Brunel University*

[53] Valyus N A 1966 *Stereoscopy* (New York: Focal)

[54] van Berkel C, Parker D W and Franklin A R 1996 Multi-view LCD *Proc. SPIE* **2653**

[55] van Berkel C and Clarke J A 1997 Characterisation and optimisation of 3D-LCD module design *Proc. SPIE* **3012**

[56] van Berkel C and Clarke J 2000 Autostereoscopic display apparatus *US Patent* 6 064 424

[57] van Berkel C and Parker D 2000 Autostereoscopic display apparatus *US Patent* 6 118 584

[58] Wartell Z, Hodges L F and Ribarsky W 1999 Balancing fusion, image depth and distortion in stereoscopic head tracked displays *Computer Graphics, Proc. ACM Siggraph99*

[59] Winnek D F 1968 Composite stereography *US Patent* 3 409 351 (issued Nov. 1968)

[60] Wheatstone C 1838 Contributions to the physiology of vision I: on some remarkable and hitherto unobserved phenomena of vision *Phil. Trans. R. Soc. (Biol.)* **18** 371–395

[61] Woodgate G, Ezra D, Harrold J, Holliman N, Jones G and Moseley R 1997 Observer tracking autostereoscopic 3D display systems *Proc. SPIE* **3012**

[62] Woodgate G, Harrold J, Jacobs M, Moseley R and Ezra D 2000 Flat panel autostereoscopic displays—characterisation and enhancement *Proc. SPIE* **3957**

[63] Woodgate G, Moseley R, Ezra D and Holliman N 2000 Autosterescopic display *US Patent* 6 055 013 (priority Feb. 1997)

[64] Woods A, Docherty T and Koch R 1993 Image distortions in stereoscopic video systems *Proc. SPIE* **1915** 36–48

[65] Yeh Y and Silverstein L D 1990 Limits of fusion and depth judgement in stereoscopic colour displays *Human Factors* **32** 45–60

[66] Yeh Y 1993 Visual and perceptual issues in stereoscopic colour displays *Stereo Computer Graphics and Other True 3D Technologies*, ed D McAllister (Princeton, NJ: Princeton University Press)

C2.7
Optical scanning and printing

Ron Gibbs

C2.7.1 Introduction

Optical scanning is familiar to the general public through such widespread applications as supermarket barcode scanners, desktop colour scanners, laser lightshows and desktop laser printers. However, scanning technology is also applied to commercial printing processes, thermal imaging, medical diagnostic equipment, biochemical analysis, and quality control of such diverse items as sheet steel, drawn wire, windshield glass, semiconductor wafers and rice—to name just a few of the many applications.

Optical scanning in electro-optical systems began to be developed for pre-press commercial print processes in the 1940s, and for airborne mapping and reconnaissance in the 1950s, but the field really took off after the invention and commercialization of the laser in the 1960s. Since then the field has developed rapidly, with military investment in forward-looking infrared (FLIR) scanners, and a steady stream of new commercial applications.

Scanning systems can be broadly divided into input and output systems. Input systems acquire electronic data in one, two or three dimensions from physical objects, by scanning a detector across the object. Output systems create an image from the electronic data by scanning a modulated light beam (usually a laser) across a light-sensitive medium. Input systems can be further subdivided into remote/local sensing and passive/active scanning. Although the various types of system have often developed separately, and in some cases are described by different terminologies, there is a great deal of overlap and common ground in the technology used.

Scanning technology is very varied, both in the components employed and in their configuration into optical systems, and no treatment of the technology can hope to be complete. Typically, to meet a given requirement, many possible solutions exist, and comparing potential solutions can involve analysis of a complex trade-off of optical, mechanical, electronic and software issues. In addition, non-scanning solutions may also be possible, especially those involving detector and light source arrays. In some instances, electronically scanned arrays can be considered to be part of scanning technology, although analysis of the system in this way is more meaningful when the electronic scan is combined with optomechanical scanning.

This chapter is organized as follows: An overview of the most commonly encountered scanning configurations is followed by a description of the most important scanning system performance requirements. Then follow descriptions of scanning system deflector and lens components, and finally more detailed descriptions of a few examples of practical input and output scanning systems. An attempt has been made to generalize, as far as possible, although inevitably many detailed issues are specific to a particular category of scanning system.

C2.7.2 Scanning system configurations

Scanning systems can take many forms, and it is helpful to group them together in related fundamental configurations. There are several commonly used ways of classifying scanning systems, which are used in different circumstances, although none of them is perfect in the sense of including all possible types of scanning system.

C2.7.2.1 Classification by deflector position

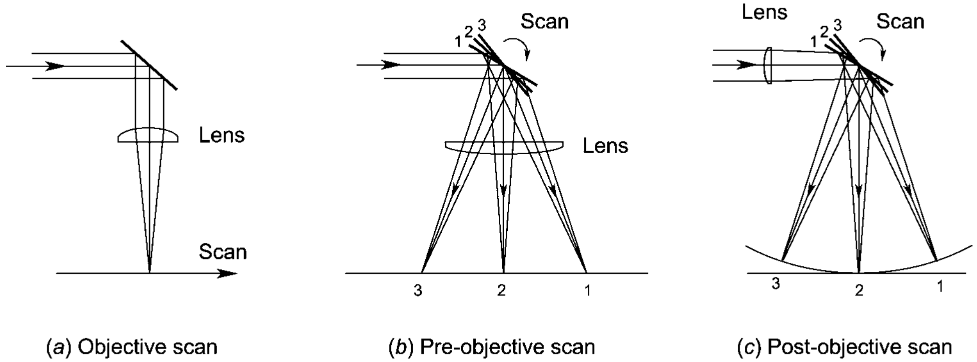
The most commonly applied classification, due to Beiser [1], is based on the relative positions of the scanning deflector and the objective (focusing lens) subsystems that make up the basic optical configuration (figure C2.7.1).

Objective scanning is defined as the coincidence of the deflector and objective, and it implies translation of either the complete optical system or the scanned object. Examples of scanning systems employing this classification include pushbroom (linear translation of a linear array), *XY* (orthogonal linear translation stages) and external drum scanning (see section C2.7.2.2). The scanning speed is limited by the mechanical inertia of large structures, which must move at the scanning speed.

Since this particularly affects acceleration and deceleration, objective scanning systems are either slow or continuous, as in the cases of external drum scanners (see section C2.7.2.2) or airborne or satellite pushbroom systems, where the aircraft flight over the ground provides the scan motion. To overcome this limitation, configurations based on angular deflectors, giving an optical lever advantage, must be used.

In *pre-objective scanning*, the deflector precedes the lens in the optical path. This usually requires a complex lens, as it must operate over the range of off-axis field angles defined by the angular scan range. However, this is the only scanning configuration that can generate a fast-scanned flat image plane via a suitably designed flat-field lens (see section C2.7.5.3), which is the usual use for this configuration. Either a single-axis deflector producing a straight-line scan, or a two-axis deflector to scan a flat plane can be used.

In *post-objective scanning*, the deflector follows the objective lens in the optical path. This allows relatively simple on-axis optics, but leads to a curved field, which can be circular if there is no separation between the rotation axis and the deflection point. This curvature can be acceptable, provided that the system either has a large depth of focus or is imaging a curved (e.g. cylindrical) surface. These curved bed systems are described further in section C2.7.2.2.



(*a*) Objective scan (*b*) Pre-objective scan (*c*) Post-objective scan

Figure C2.7.1. Scan configurations by deflector position.

Additional optics (usually based on large aspheric mirrors) can be inserted between the deflector and the image to flatten the image field. These are often described as post-objective field flatteners, but the resulting systems have more of the characteristics of pre-objective scanners.

This classification system deals with individual scanning mechanisms for a single scan direction. In two-dimensional scanning systems, two orthogonal scanning mechanisms are combined. In the case of a raster scan pattern, fast scan and slow scan axes are defined. Often, the fast scan mechanism uses angular deflectors in a post- or pre-objective configuration, and the slow scan mechanism uses objective scanning.

C2.7.2.2 Classification by image geometry

Particularly in applications that involve flexible object or image media, it is useful to classify scanning systems by the (input or output) imaging medium surface form, which can be either flat or curved when mounted for scanning. The medium could either be the surface of a solid object (e.g. a cylinder) or a thin, flexible sheet of material, which is constrained to take the shape of a mounting surface, or bed. This classification system deals with complete scanning system geometries.

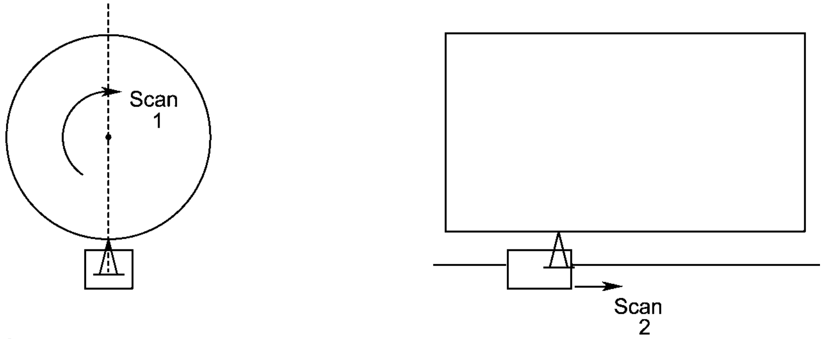
A *flatbed* scanner usually has the benefits of convenient material geometric form and simple handling mechanisms, together with the possibility of a compact, lightweight system. Usually, a pre-objective scanner or linear array is used, often requiring a complex lens design (see section C2.7.5). It is possible to flatten the curved field of a post-objective system by using dynamic focus of the objective, but this is limited to slow scanning speeds because of the mechanical inertia of the lens. A closely related form is the capstan scanner, which combines a straight-line scanner with a transport mechanism for the input or output sheet that moves it across the scan line using motorized rollers. Typically, this winds a photographic film from a roll to a take-up roll.

Non-flatbed scanners are based on cylindrical surfaces, either convex or concave. Convex cylinder scanners are called *external drum*, as the scanning optics is external to the cylinder surface. In these, the optical system is usually stationary and the cylinder rotates (figure C2.7.2). In this configuration, there is great system flexibility for the optical design, which can vary from simple (single-beam, fixed resolution) to complex (multiple beams, variable resolution). In addition, the (mechanical) scanning system is separate from the optical system, so it is relatively easy to convert an existing system by changing the optics module, or to change the scanned image format by changing the size of the cylinder. As the focal length of the optical system can be small, very high resolution and good image quality can be achieved, but the scan speed is limited by the very high rotational inertia of the large cylinder drum. This causes problems for acceleration and deceleration, as well as safety and stability concerns that limit rotation speeds to about 2000 rev/min.

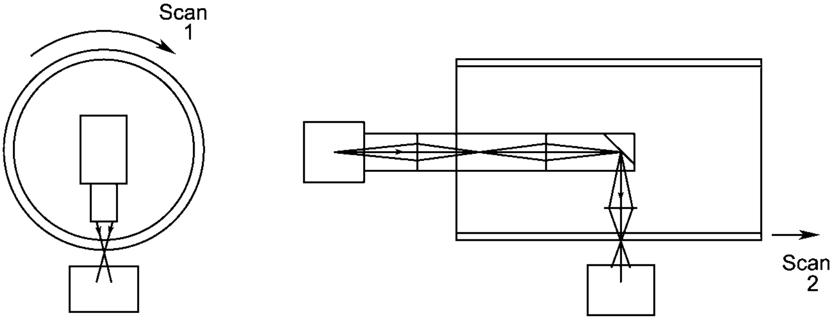
An alternative form of external drum scanner uses a line scanner to generate a line along the cylinder, which rotates only once for each complete scan of the drum surface. This is the configuration for typical xerographic laser scanners.

Concave cylinder surface scanning systems are based on post-objective scanners, which have simple on-axis lens optics, and which can take one of the two forms, as described above. When the deflector rotation axis is parallel to, and collinear with, the optical axis, an *internal drum* scanner results (figure C2.7.3). This has the advantage that as the deflector is small, very high rotation rates can be attained. In addition, a large scanning angle range (up to 270°) is easily achieved, and fairly high resolution scanning is possible over a fairly large area. The image surface is stationary, which simplifies and speeds up material changeover between scans.

Two forms of internal drum scanner optical layouts may be identified. If the laser and optical systems are small, the complete optics may be mounted on the linear traverse slide, in a carriage-mounted



(*a*) Solid drum
Output, reflective input



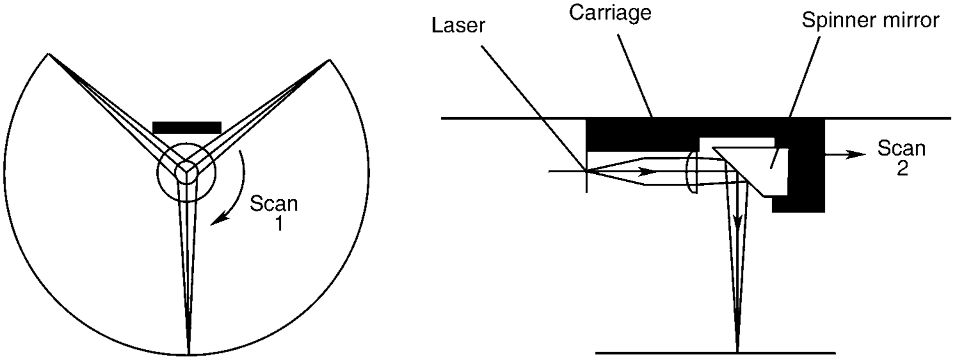
(*b*) Transparent drum
Transparency input

Figure C2.7.2. External drum configurations.

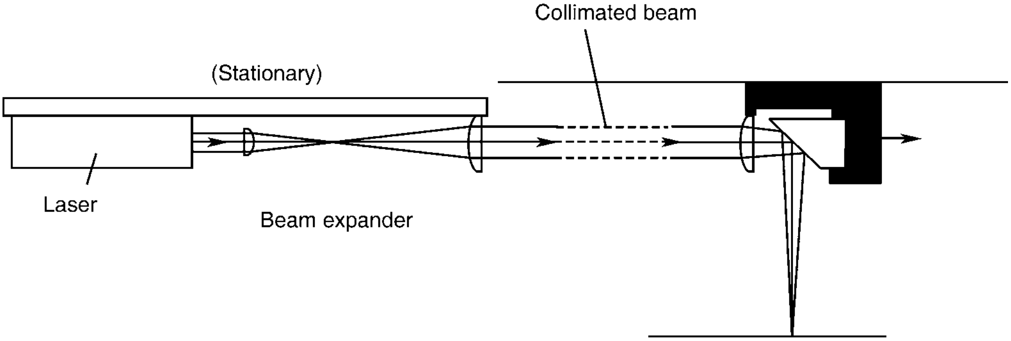
configuration. When the optics is too large or heavy, only the final focusing lens and the spinner mirror is mounted on the carriage, and the preceding optics generates a collimated beam collinear with the cylinder axis. The beam must be sufficiently large that the beam divergence is small over the carriage traverse length. This is known as a beam-riding configuration.

Multiple-beam internal drum systems are difficult to realize, because of the image rotation along the scan line inherent in the scan geometry. A half speed contra-rotating dove prism before the spinner mirror can correct this [2], although the speed control requirements are very demanding. By using a two-axis acousto-optic deflector, derotation of up to three beams is possible [32].

A post-objective scan configuration with the deflector axis perpendicular to the optical axis (such as in figure C2.7.1*c*) is classed as *curved bed*. Compared with the internal drum configuration, only a small angular scan is possible, but the deflection angle is double the mirror rotation angle, resulting in a faster scan. Multiple-beam scanning is possible, as there is no rotation on the scanned beam axis. For faster scanning, a prismatic polygon mirror may be used in place of a single-facet mirror. However, in this case, the separation between the rotation axis and mirror surface results in distortion of the image surface curvature and the scan linearity.



(*a*) Carriage-mounted optics



(*b*) Beam-riding optics

Figure C2.7.3. Internal drum configurations.

C2.7.2.3 Classification by scan pattern

The main division is between raster and vector scanning systems. Raster scanning repetitively scans a beam in a regular pattern of adjoining lines over a rectangular area, like a TV display. Raster scanning systems are characterized by:

- fixed scan speeds, often set by a system electronic clock;
- orthogonal scan mechanisms—one being a fast line scan and the other being a much slower page scan.

For multiple-beam or multiple-detector systems, a further sub-classification can be applied, depending on the method of overlapping the beams. In a swathe scanning system, each set of multiple beams forms an overlapping scan pattern or swathe, as in a pushbroom scanner. To scan a larger width, multiple scans are made with the first beam of each swathe adjoining the last beam of the preceding swathe. Some systems allow adjacent elements of detectors (e.g. CCD) or light sources (e.g. LED), but when some physical separation of elements is necessary, the array must be angled with respect to the scan direction, to allow overlap of adjacent elements. In this case, a timing correction must be made to the electronic signals to or from the array elements.

If an array of n elements has a spacing between elements of $(n+1)p$, where p is the required separation between adjacent scan lines, an interconnecting scan pattern will eventually result if each pass of the array is advanced by a distance np in the slow scan direction. This is known as an interlaced scan.

In page width segmentation, in a single pass of the scanner the beams are equispaced across the image width. On each pass of the scanner, the beams are moved across by one beam width so that each beam adjoins the corresponding beam from the previous scan, until it reaches the starting position of its following neighbour.

Random-access and vector scanning systems require agile, low-inertia deflectors to move a scanning spot along arbitrary paths within a scanning area. Typically, orthogonal galvanometer deflectors or electro-optical deflectors are used for very high speed scanning applications. In vector scanning, the two deflectors are synchronized to move the spot at constant speed, to maintain a constant exposure. In random access, the spot is quickly switched to a new direction, requiring sophisticated control to minimize the switching time without causing oscillation.

Complex scan patterns, similar to Lissajous figures, may be generated by simple scanning systems comprising of two fixed-speed oscillating deflectors. The deflectors could be resonant galvanometer scanners, or rotating transmissive glass wedges known as Risley prisms. Different combinations of the relative speed and phase of the two deflectors produce a wide range of scan patterns [3].

C2.7.3 Scanning system performance specification

In this section, the performance parameters that are relevant to scanning systems are examined. Scanning system performance requirements vary widely in terms of resolution, speed and quality, and a thorough analysis of these requirements is essential before any design is attempted.

C2.7.3.1 Resolution

Resolution of a scanning system can be expressed in several different ways. The addressability of a digital system is determined by the sampling frequency of acquisition or output of pixel data points, and is usually measured in units of dots per inch (dpi). The image is usually sampled at finer intervals than the optical resolution of the system. Hence, addressability can give a misleadingly high value for the system resolution if not interpreted with care.

The spot size (of the scanning laser beam, or of the detector element image) is a measure of the optical resolution of the system. The cross-scan (perpendicular to the scan direction) resolution is well characterized by spot size, but the in-scan (in the direction of the scan direction) resolution of a flying spot scanner is degraded by 'smearing' of the scanned spot as a result of its motion. In the cross-scan direction, the ratio of the spot size to the scan line pitch is significant, as this determines the intensity ripple in this direction, and hence the visibility of scan lines.

In a laser scanning system, the beam usually has a gaussian intensity profile, and the spot size is determined at either the 50% (referred to as full-width, half-maximum (FWHM)) or $1/e^2$ (13.5%) level. In some systems, the spot is elliptical in shape, with the size optimized for both in-scan and cross-scan resolutions, but where a round beam is used, the spot size is a compromise between the two requirements.

For an angular scanned system, the concept of angular scanned resolution is useful. Angular scanned resolution is defined in terms of the number of resolved spots per scan line, N.

$$N = \theta/\Delta\theta \tag{C2.7.1}$$

where θ is the maximum scan angle and $\Delta\theta$ the scan angle increment corresponding to the resolution limit of the system. For an optical aperture width D, the resolution is determined by the diffraction limit

$$\Delta\theta = a\lambda/D \tag{C2.7.2}$$

where λ is the wavelength and a a constant that depends on the aperture shape, beam uniformity and resolution criterion used. (For the common case of an untruncated gaussian beam defined by the FWHM spot size, $a = 0.75$.) Hence, the fundamental scanning equation [31] is obtained

$$N = \theta D/a\lambda. \tag{C2.7.3}$$

A consequence of this equation is that by increasing the scan angle with additional optics such as an afocal beam expander does not improve the number of resolved spots per scan line, as θD is an optical invariant.

The calculated modulation transfer function (MTF) can be very useful when applied to a two-dimensional raster scanning system used for imaging [4]. Although MTF is only strictly applicable to linear systems, and some elements of scanning systems (especially electronic and software image processing) can be highly non-linear, the calculations can still provide valuable insights if treated with caution.

In-scan and cross-scan MTFs are different, because of the time-dependent factors that affect in-scan resolution, which are linearly related to spatial frequency response by the scan speed.

For an input system, the model must include:

- lens performance, or aperture, if diffraction limited; calculated or measured;
- detector element size and shape; calculated from the Fourier transform of the detector dimensions;
- scan motion; calculated from velocity as a sinc function [4].

In addition, electronic responses of scanner system components must be included:

- detector frequency response;
- detector amplifier frequency response;
- software processing of the output data (frequency response is linearly related to MTF).

For output systems, the key optical factors in the MTF calculation for the cross-scan direction are

- optical spot size and intensity profile—calculated from the Fourier transform;
- output medium—measured (related to the microscopic grain structure of the active material).

Additional system factors for the in-scan direction are

- spot motion;
- modulator frequency response, due to the optical modulation device and the electronic amplifier response.

Measurement of scanning system MTF is difficult for both input and output systems, due to problems in generating sinusoidal spatial frequencies and because of the complex sampled nature of the image. In practice, spatial resolution is usually assessed by subjective methods, using test images and resolution charts.

C2.7.3.2 Scan duty cycle

Duty cycle is an important measure of the efficiency of the scanning system. It is usually defined for a line scanning subsystem as the proportion of time that the system is actually performing input or output of scanned data during the scanning operation. It arises because every scanning mechanism has a 'dead' time, which arises from the scanning geometry.

Examples of how dead time arises for specific scanning mechanisms are:

- Galvanometer: acceleration/deceleration at start/end of linear angular scan range, and flyback time if the scan is unidirectional.
- Polygon: transit time of the beam across corners of the polygon mirror.
- Internal drum: obscuration of the deflected beam by the spinner carriage.
- Acousto-optic deflector: acoustic fill time.

The duty cycle has a significant effect on the system electronic bandwidth requirement, and on the radiometry. The electronic bandwidth of the system must be higher than the average data rate due to this inefficiency in the optical system. For example, the optical modulation rate f_{m} of an output system is related to the average data rate f_{d} by

$$f_{\mathrm{m}} = \frac{f_{\mathrm{d}}}{\eta}. \tag{C2.7.4}$$

The effect on radiometry is that higher optical power is required to illuminate or expose the scanned object than if the scan were perfectly efficient. In some cases, this can limit the achievable scanning speed of the system.

C2.7.3.3 Radiometry, input systems

The critical radiometric performance measure for input systems is the dynamic range of the system. This is defined as the ratio between the system's white and black levels I_{w} and I_{b}, which are the maximum and minimum possible detector signals.

The white level can depend on the following factors:

- the brightness of the optical source, including its condenser optics;
- the optical properties (e.g. reflectivity or transmissivity) of the scanned object;
- the efficiency (e.g. numerical aperture, transmission) of the detector optical system;
- the size, responsivity and saturation level of the detector element.

The dynamic range may be expressed as a simple ratio, but is more commonly expressed in density units

$$D = \log_{10}\left(\frac{I_{\mathrm{w}}}{I_{\mathrm{b}}}\right) \tag{C2.7.5}$$

or as an integer number of bit levels

$$N_{\mathrm{D}} \le \frac{\ln(I_{\mathrm{w}}/I_{\mathrm{b}})}{\ln 2}. \tag{C2.7.6}$$

Optical density is often a more relevant measure of signal, because of the logarithmic response of many physical processes, including photographic emulsions and the eye. N_{D}, as defined here, is the maximum number of physically significant bit levels that can be achieved in a digitized output.

C2.7.3.4 Radiometry, output systems

For exposure of an image on an area A, the optical power of the beam is given by:

$$P = \frac{ks}{\eta_{\mathrm{opt}}}\left(\frac{A}{\eta_{\mathrm{L}}\tau}\right) \tag{C2.7.7}$$

where s is the sensitivity of the exposure material, η_{L} the scan duty cycle, η_{opt} the optical throughput efficiency and τ the total exposure time.

For most scanning systems, the material sensitivity must be valid for very short exposure times. This can differ significantly from manufacturers' measured values at shorter exposures because of reciprocity failure. For photographic materials, the sensitivity usually increases at shorter exposures, whereas for thermal materials it decreases.

The value of the constant k in the equation depends on the beam profile and the required overlap between adjacent scan lines. When $k = 1$, a threshold exposure results, when only a thin line at the centre of the scan line is exposed. A typical value for a practical system is approximately 2.

Note that, for a fixed scanner speed in terms of the number of scan lines per second, as the system resolution increases, the exposure time increases. Hence, the power requirement also increases.

C2.7.4 Scanning deflectors

Angular scanning systems use deflector subsystems to deflect the optical axis of a beam from a point within the scanning system in the form of a rotation about a scanning axis. A variety of deflection subsystems have been developed to meet the needs of different scanning applications. The deflector subsystem, with or without auxiliary focusing optics, is itself often described as a scanner by manufacturers.

Distinction may be made between high-inertia deflectors, which can only be used for continuous, constant speed scanning, and low-inertia deflectors, whose deflection angle can be varied quickly. The latter category often has a limited angular deflection range, which limits its application to low-resolution systems. Low-inertia deflectors include fast-steering, or tip-tilt mirrors, which are two-dimensional beam steering devices primarily used for line-of-sight stabilization or tracking applications.

C2.7.4.1 Galvanometer mirror

This important class of deflector employs the electromagnetic effect of a permanent magnet and the magnetic field created by an electric current in stator coils to produce a rotary torque, which drives a spindle through a limited angular range [5]. A mirror is mounted on the end of the spindle to deflect the optical beam. The mirror is a significant contributor to the mechanical inertia of the rotor, which limits the maximum speed and/or angular range of the scanner. Angular range is typically no more than 40° (80° optical deflection) for low-accuracy applications, and less than 20° for high-end applications.

The rotor can be configured as moving iron, moving magnet or moving coil [33]. Originally, galvanometer scanners were based on moving iron designs [6], but modern designs are based on moving coil or moving magnet designs. Moving coil designs have relatively high inertia, but high and stable torque, and are typically used for high-accuracy applications with large mirrors (greater than 30 mm aperture). Moving magnet galvanometer scanners, aided by recent developments in high-energy-density rare-earth magnetic materials, have the highest torque-to-inertia ratio, and hence the highest speeds. The angular deflection speed and position are controlled by sophisticated closed-loop servo electronics for high-accuracy applications, angular position being accurately measured using accurate, capacitive or optical sensors with fast response times. Resolution of the order of one microradian can be achieved.

For two-dimensional scanning, two deflectors are often used, mounted close together with their axes at right angles. In this configuration, the second mirror must be larger than the first, to accommodate the beam movement. Because of the separation of the scan axes, the scan pattern is geometrically distorted, but this can be computed [7] and compensated by software correction of the drive signals to the galvanometers.

Galvanometer scanners are most often used in one of two modes: vector scanning or regular periodic form. In vector scanning, the two deflectors of a two-dimensional scanner are controlled in synchronism to move the optical beam along a programmed path in response to a required pattern (e.g. alphanumerical characters), usually at a constant spot speed. For regular (e.g. raster) scanning, the deflector is driven so that the deflection is linear with time over a large proportion of the scan period.

For a high duty cycle a sinusoidal variation is inadequate, and the deflector angle is controlled to a sawtooth form, where the beam that is scanned at linear velocity in one direction is, quickly returned to the starting point. The scan must be unidirectional because of the slight ellipticity in the scan line due to bearing hysteresis.

For very high fixed scanning rates, resonant scanners have been developed. These use flexure, torsion or taut-band mounts to enable very high scan speeds of up to 8 kHz or more. Although these deflectors have low inertia, the scan speed and sinusoidal oscillation are fixed, so electronic signal timing correction is required to obtain an undistorted scan with reasonable scan duty cycle.

C2.7.4.2 Polygon mirror

Deflectors based on continuously rotating spindles are very commonly used in scanning systems, and may have either a single mirror facet (monogon) or many facets (polygon).

These deflectors have much higher mechanical inertia than galvanometer scanners, but can rotate at very high speeds, thanks to the development of precision air bearings. These bearings have largely superseded ball bearings for high-end applications, having much higher speeds, much lower vibration and longer lifetime. Bearings can be hydrostatic, i.e. externally pressurized, requiring a source of compressed air. However, many bearings nowadays are hydrodynamic (or self-acting), requiring no external air supply. These require much tighter manufacturing tolerances, but modern manufacturing techniques, coupled with developments in bearing design (especially materials and surface finish), make these a practical choice for most high-speed applications.

At very high rotation speeds (>30 000 rpm), air resistance (windage) and turbulent flow of air at the mirror edges limit the achievable speed, and very high-performance scanners are often enclosed in partial vacuum or in a helium atmosphere, which has much lower viscosity than air [8].

Polygon mirrors may be pyramidal or (more commonly) prismatic. Prismatic polygons have their facet surfaces parallel to the rotation axis, while the facets of pyramidal polygons are at an angle (usually 45°) the rotation axis. Prismatic mirrors deflect the optical beam by twice the mechanical rotation, whereas pyramidal mirrors deflect it by the rotation angle.

Prismatic polygons are most commonly used in a configuration that has the input and output beams in the same plane, perpendicular to the polygon rotation axis (figure C2.7.4). This configuration avoids

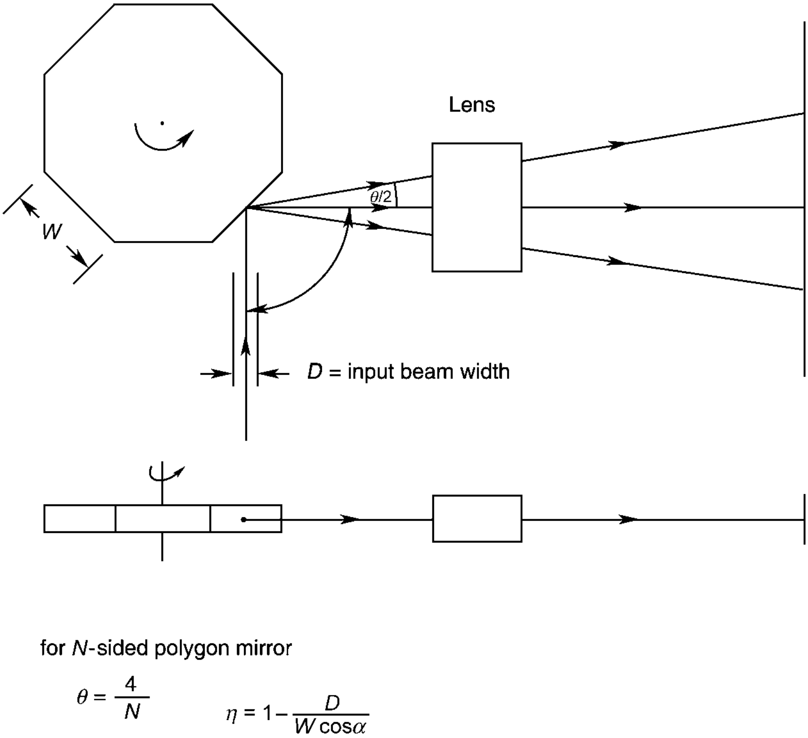

Lens
$\theta/2$
W
D = input beam width
for N-sided polygon mirror
$\theta = \frac{4}{N}$
$\eta = 1 - \frac{D}{W \cos\alpha}$

Figure C2.7.4. Prismatic polygon scanner configuration.

scan line bow. Careful calculation is required to design the polygon dimensions to achieve the required system performance with the minimum size [9, 10].

Active facet tracking can be used to further reduce the size of the polygon facets. In this technique, a subsidiary small-angle deflector is inserted before the polygon. This high-bandwidth deflector is programmed in such a way that it always deflects the input beam to the centre of the active polygon facet.

Pyramidal mirrors (figure C2.7.5) are more expensive to fabricate, but can achieve very high scan speeds in a compact configuration. They are often used overfilled, i.e. with an input beam larger than the mirror facet. This results in a scanned spot of constant power and a high scan duty cycle, but the optical transmission efficiency is poor. It is possible to design a coarse two-dimensional scan system by varying the pyramidal angle between adjacent facets to produce an angular step between successive scan lines.

The single-facet (monogon) versions of prismatic and pyramidal deflectors are shown in figure C2.7.6. The pyramidal type is the more commonly used, usually in a post-objective internal drum scanner, and many innovations have been applied to increase the speed. Optically transparent enclosures have been developed to overcome windage limitations [11], and special mirror shapes have been developed to reduce dynamic distortions of the cantilevered mirror surface [12].

C2.7.4.3 Holographic

Holographic deflectors [34] are based on rotating (flat or curved) surfaces containing fine diffraction gratings, which can be either linear gratings to deflect collimated beams, or more complex curved

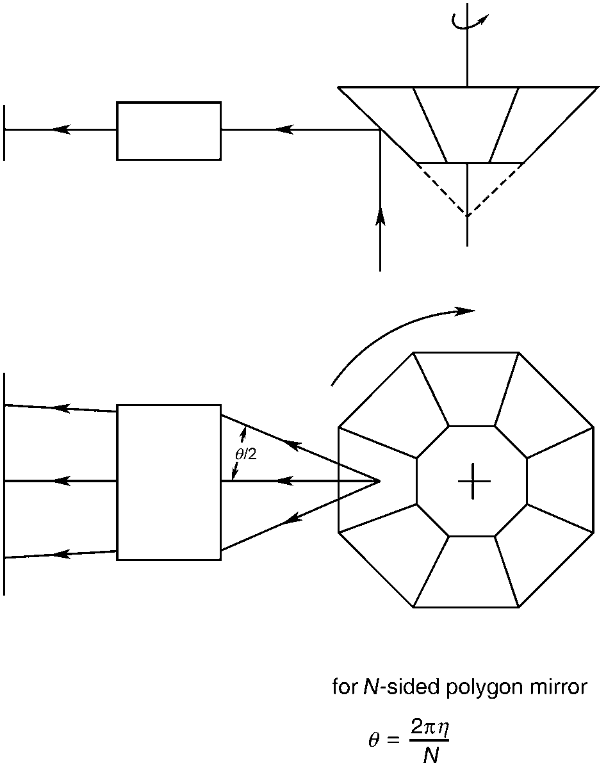


Figure C2.7.5. Pyramidal polygon scanner configuration.

structures that have optical power and act as lenses. Thick blazed transmissive gratings can have high transmission efficiency in a single diffraction order for a single wavelength. A single surface that combines several holographic elements is referred to as a hologon (from *holo*graphic poly*gon*).

The disc hologon is a form in which hologram elements are arranged radially around a flat circular disc, which rotates about its axis (figure C2.7.7), and performs a similar function to a prismatic polygon [13]. The figure shows a hologon with optical power, but linear gratings and a separate lens are often employed in a pre-objective scanning configuration. In this case

$$\sin\theta_\mathrm{i} + \sin\theta_\mathrm{d} = \frac{\lambda}{t} \tag{C2.7.8}$$

where t is the grating period.

If the hologon is tilted by a small angle $\mathrm{d}\alpha$, the change in diffracted angle is given by [14]

$$\mathrm{d}\theta_\mathrm{d} = \left[1 - \frac{\cos(\theta_\mathrm{i} + \mathrm{d}\alpha)}{\cos(\theta_\mathrm{d} - \mathrm{d}\alpha)}\right]\mathrm{d}\alpha \tag{C2.7.9}$$

from which it can be shown that sensitivity of the output angle to wobble in the hologon tilt is minimized when $\theta_\mathrm{i} = \theta_\mathrm{d}$, which is the Bragg diffraction condition.

An alternative form of holofacet deflector is based on holograms arranged around a cylindrical surface that rotates about the cylinder axis (figure C2.7.8), and is analogous to a pyramidal polygon.

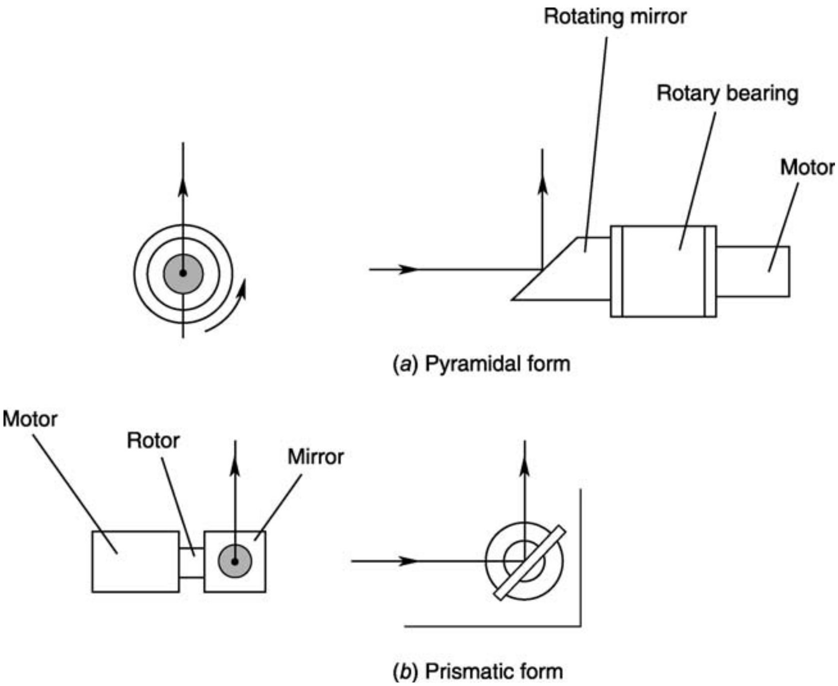

Rotating mirror
Rotary bearing
Motor
(*a*) Pyramidal form
Motor
Rotor
Mirror
(*b*) Prismatic form

Figure C2.7.6. Single-facet (monogon) deflectors.

Holographic deflectors have several potentially attractive fundamental features, compared with polygon mirrors:

- Thin lightweight elements, with improved aerodynamic shape (elimination of sharp edges in the rotating part), therefore potential for higher-speed rotation.

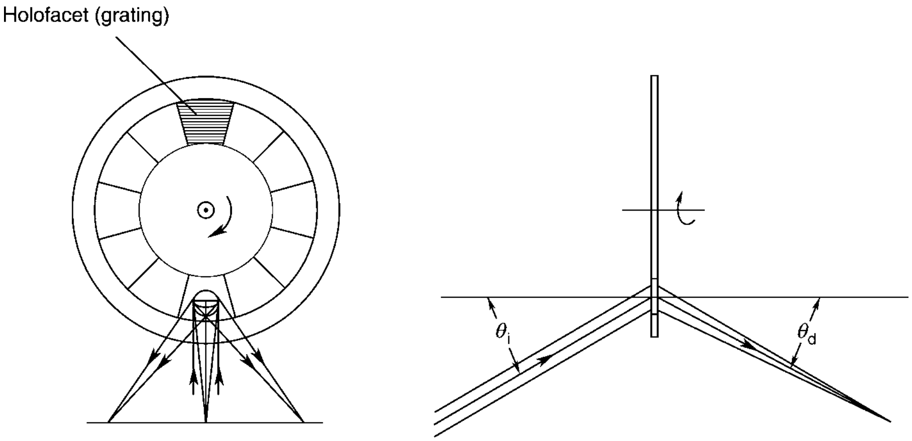
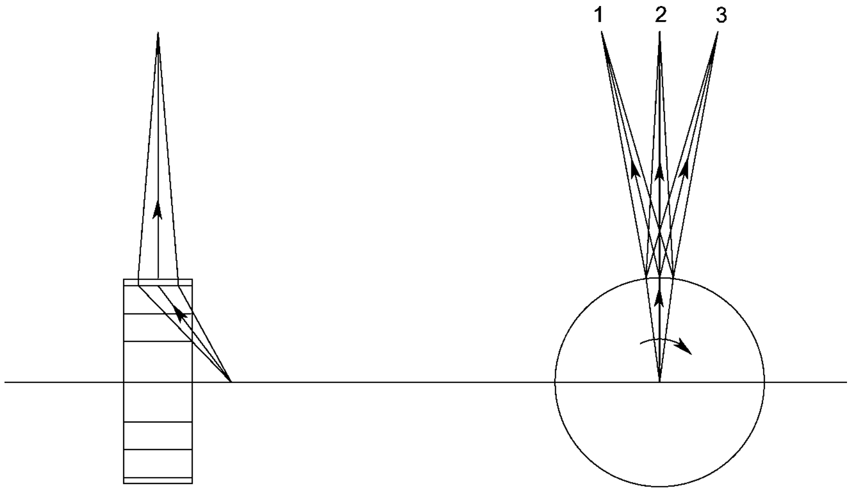
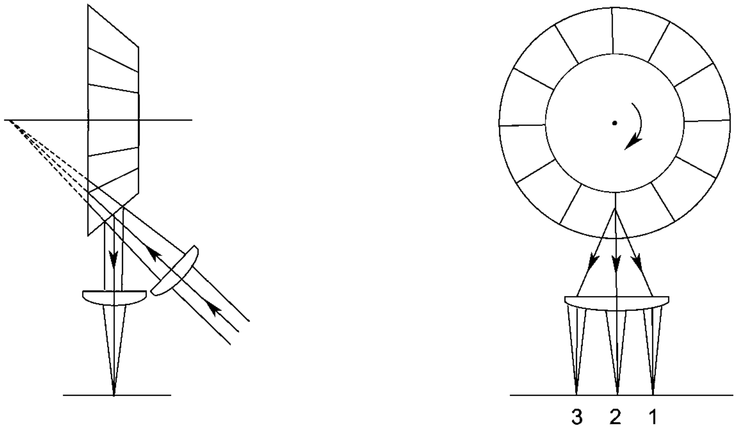

Holofacet (grating)
θ_i
θ_d

Figure C2.7.7. Disc hologon.



(*a*) Transmissive



(*b*) Reflective

Figure C2.7.8. Cylinder hologons.

- Reduced centripetal deformations and/or reduced sensitivity to such deformations, therefore improved accuracy at high rotation speeds.
- Reduced wobble sensitivity (at Bragg angle), therefore potential elimination of anamorphic deflection error correction optics.

In addition, they offer increased design flexibility to enable compact configurations, and are potentially more accurate and/or economic to be manufactured in quantity.

However, they have some limitations:

- Wavelength sensitivity of deflection angle, which must be corrected by auxiliary diffractive optics when used with typical laser diodes [15].

- Specialized and technically demanding design and fabrication requirements, few suppliers having the necessary skills and facilities, and typically a lengthy, expensive development for a new system.
- The most useful configurations are radially asymmetric, leading to complex and difficult optical system design issues; in particular [14], careful consideration may need to be given to:
    - Scan bow and scan linearity; arising from complex, asymmetric scan geometry.
    - Radiometric uniformity; diffraction efficiency varies with angle of incidence.
    - Polarization: diffraction efficiency depends on the relative orientations of the input beam polarization and the grating, unless the beam is circularly polarized.

A single-facet holographic deflector, comprising a single linear grating, has been successfully used as a monogon deflector [14]. This is designed to have a diffraction angle of approximately 90°. The grating is mounted at 45° to the input beam, when the wobble insensitivity is most effective, in a mount that rotates about the beam axis. The rotating mount can include a lens after the grating, for use in internal drum configuration, or the deflector can be used with a fixed flat-field lens in a pre-objective scan configuration.

C2.7.4.4 Optoelectronic

The term optoelectronic is used here to denote any deflector that has no mechanical moving parts, and whose operation depends on the interaction of an electrical signal with the optical properties of the deflector material to produce an angular deflection. These deflectors can have effectively zero inertia, the response being limited only by the speed of the drive electronics, making them ideally suited to broadband position control and ultra-high speed scanning. However, in practice, they are only rarely viable alternatives to mechanical scanning mechanisms, due to their limited angular deflection range and optical aperture, and hence poor angular resolution, typically much less than 1000 spots per scan line. Their application is therefore generally limited to low-resolution very high-speed imaging and active error correction subsystems.

Electro-optic (EO) deflectors use the electro-optic Pockels effect of some crystal materials to deflect a linearly polarized beam by an angle proportional to an applied voltage. Electro-optic coefficients of even the best crystal materials have low values, so angular deflection is small. Because of the birefringence and thermal sensitivity of available EO materials, sophisticated designs are required, often passively compensating the undesirable effects of temperature and mechanical stress using additional crystals.

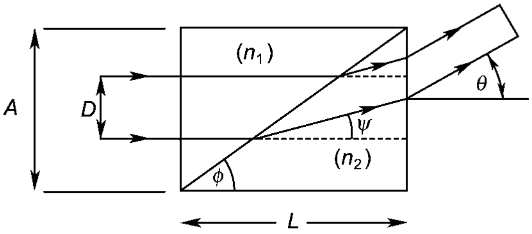
One commercial implementation produces a uniform refractive index gradient across the crystal using a quadrupole array of electrodes [16]. An alternative form of EO deflector works by inducing a refractive index change in a prism (figure C2.7.9).

The deflection angle may be expressed as

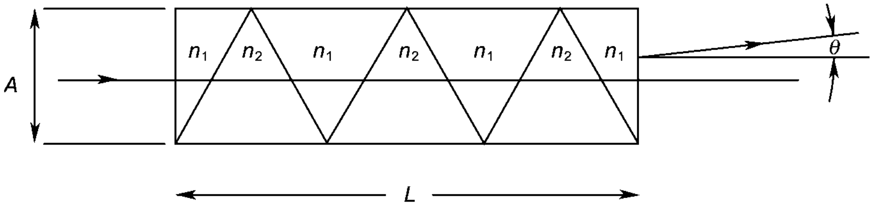
$$\theta = \frac{\Delta n}{n}\frac{L}{D} \tag{C2.7.10}$$

where $\Delta n = (n_1 - n_2)$ and n is the refractive index at the final air interface. For an EO material, the change in refractive index with applied voltage V_z is given by

$$\Delta n = 2n^3 r_{ij}\frac{V_z}{t} \tag{C2.7.11}$$



(a) Single prism



(b) Multiple prism

Figure C2.7.9. Prism-type electro-optic deflector.

where r_{ij} is the material electro-optic coefficient and t the thickness of the cell in the z direction (perpendicular to the paper).

The deflector performance is improved by fabricating the deflector in the form of several sequential prisms of alternating polarity, increasing the length of the cell while allowing practically achievable crystal sizes. This has been implemented using semiconductor-manufacturing technology, growing $LiTaO_3$ crystal prisms using photolithography [17].

Acousto-optic (AO) deflectors use diffraction in a crystal material, produced by pressure waves, to deflect the beam by an amount dependent on the applied frequency f. The deflection results from a thick diffraction grating created within a crystal material by a radio-frequency travelling pressure wave generated by a piezo-electric acoustic transducer bonded to the side of the crystal (figure C2.7.10). The first order deflection angle is given by the diffraction equation (C2.7.8) above, where the grating period $t = f/v_a$, where v_a is the acoustic velocity of the cell material.

For high diffraction efficiency, Bragg diffraction is used, where the deflection angle of the diffracted beam at the centre of the scan line is twice the angle of incidence. Varying the frequency of the acoustic wave about the mean frequency varies the deflection angle, producing a scan.

The number of resolved spots per scan line for an acousto-optic deflector is given by [18]:

$$N \approx \frac{\tau \Delta f}{a} \tag{C2.7.12}$$

where τ is the transit time of the acoustic wave across the optical aperture and Δf the change in transducer frequency from the centre frequency.

To maximize τ, a material with a slow acoustic velocity (usually TeO_2) is used and the size of the laser beam in the acoustic wave direction is made as large as possible, although these factors increase the response time of the device. To keep the size of the crystal within practical limits, cylindrical lenses are used to shape the beam within the deflector. Practical considerations of crystal size, acoustic attenuation and deflector speed limit the beam width to about 50 mm.

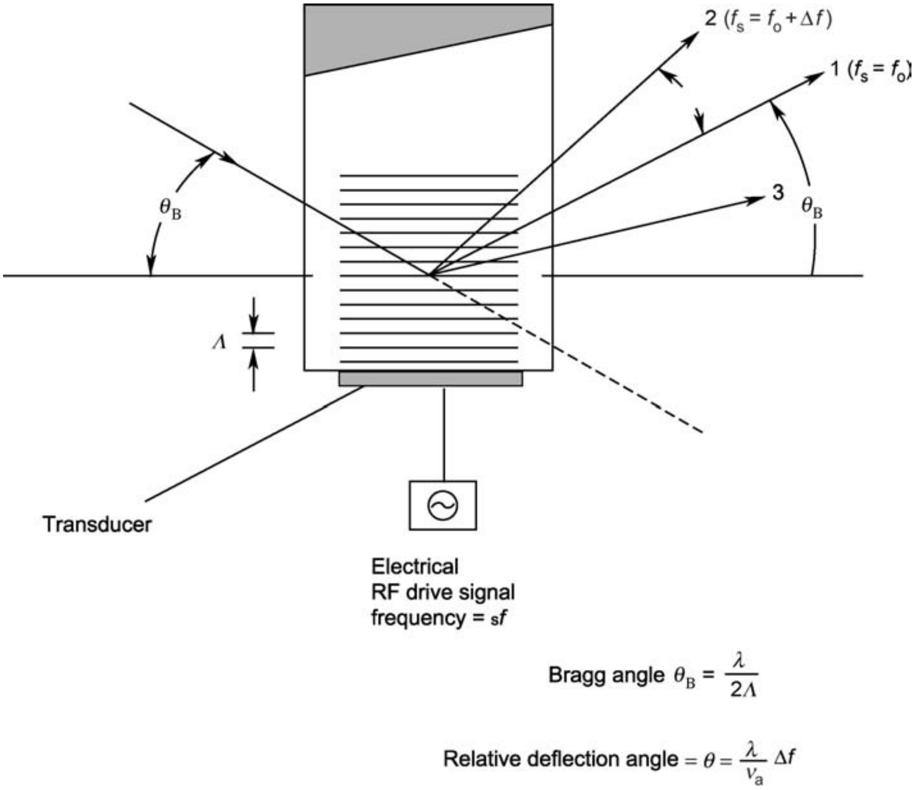

2 ($f_s = f_o + \Delta f$)
1 ($f_s = f_o$)
3
θ_B
θ_B
Λ
Transducer
Electrical
RF drive signal
frequency = sf
Bragg angle $\theta_B = \frac{\lambda}{2\Lambda}$
Relative deflection angle = $\theta = \frac{\lambda}{v_a} \Delta f$

Figure C2.7.10. Acousto-optic deflector.

To ensure that the deflected beam has constant power over the scan range, it may be necessary to use a phased-array transducer design [19] to maintain Bragg diffraction over the range of deflection angles.

During scanning, the acoustic wave frequency varies across the beam aperture. The frequency gradient produces a cylindrical lens effect with a focal length of

$$f_{\mathrm{cyl}} = \frac{v_{\mathrm{a}}\tau}{\theta_{\mathrm{T}}} = \frac{v_{\mathrm{a}}^2}{\lambda(\mathrm{d}f/\mathrm{d}t)}. \tag{C2.7.13}$$

The output beam is converging or diverging, depending on the scan direction. This can prevent bi-directional scanning, and requires the use of an external compensating cylindrical lens in the optical system.

Two-dimensional AO deflectors have been produced, using transducers on orthogonal faces of a single crystal, although practical limits to crystal size mean that the achievable angular resolution is severely limited.

Recent research has employed optoelectronic devices with high spatial resolution, such as liquid crystal or micromirror arrays, to produce deflection via phased array phase shift [20, 21] or variable-pitch diffraction gratings [22]. These devices are still in the early stages of development, and show some promise of enabling fast beam steering with lower-cost devices and drivers, but show little improvement in resolution performance over the existing acousto-optic and electro-optic technology.

C2.7.5 Scanner optics

C2.7.5.1 Reduction-type array imaging lens

In a reduction-type CCD scanner, a high-quality multi-element imaging lens images the scanned document on the detector array. This lens must be specially designed for the purpose, because a standard photographic quality camera or enlarger lens is not suitable.

The requirements that must be addressed in the design include:

- Magnification: consider both the ratio between document width and array length and the ratio between required spatial resolution (in terms of addressability) and detector element spacing. In some systems, the lens must be corrected for a range of magnifications, requiring multi-configuration optimization.
- Focal length: a compromise between mechanical layout and achievable optical performance. Short focal length increases the field angle, which makes the design more difficult.
- Spatial resolution: MTF must be specified for both sagittal and tangential directions, for the full range of field positions, and for each wavelength. The effect of sampled imaging must be considered; in particular, the Nyquist frequency acts as a limit to resolution.
- Colour correction: for colour scanners, excellent colour correction is required, often involving secondary spectrum correction and therefore careful glass type selection. Lateral colour is often a limiting aberration, which causes highly visible colour fringes on off-axis image features.
- Aperture: often determined by the radiometric requirements of the detector. The effect on diffraction-limited resolution and lens performance must also be considered.

The complexity of lens designs ranges from relatively simple Cooke triplet types to complex double-Gauss derivatives with six or more elements.

C2.7.5.2 Contact-type array scanner optics

Input scanning using a long array detector with small elements can be achieved without complex lenses, if the scan line length and resolution can be identical to those of the array. In this case, a configuration with the detector in contact with the scanned object is usually not feasible because

- Movement of the detector across the object will cause scratching.
- For a reflective object, there is no clear path for illumination of the object around the detector.
- Detector arrays are usually packaged in a hermetically sealed form, with a window and air space between the sensor surface and the object. This gap results in a loss of light collection efficiency and resolution.

The usual solution is to include a micro-optical system that images the object on the detector at unity magnification. This system can be based on either

- a linear array of glass microlenses and microprisms;
- a linear SELFOC (gradient index lens) array lens;
- an optical fibre array.

The magnification in these systems must be non-inverting, so that the images from the array lenses overlap to form a linear image at the detector array. This results in a compact, low-cost configuration, and is commonly used in hand-held and desktop scanners. Compared to reduction-type systems, the numerical aperture and the resolution of these optical systems are low, and the performance of the very long arrays is inferior to wafer-scale CCD arrays.

C2.7.5.3 Flat-field laser scanning lenses

In a pre-objective flatbed scanning system, a specially designed lens is required to focus the scanned beam to a line. A theoretically perfect thin lens produces a curved field of radius f, so a flat-field lens must be designed with the correct amount of negative field curvature to compensate the natural image curvature.

In addition, in most cases, it is highly desirable that a constant rotation rate of the scanner deflector should result in a scan spot with linear velocity at the lens focus. Thus, the linear spot deflection y at the image plane must be related to the scan angle by the relationship, known as the $f\text{–}\theta$ characteristic

$$y = f\theta \tag{C2.7.14}$$

where f is the lens focal length and θ the scan angle, so that

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f\frac{\mathrm{d}\theta}{\mathrm{d}t}. \tag{C2.7.15}$$

A distortion-free lens has the characteristic

$$y = \tan\theta \tag{C2.7.16}$$

so the lens must additionally be designed with the correct amount of barrel distortion to produce the desired $f\text{–}\theta$ characteristic. Such a lens is therefore referred to as an *$f\text{–}\theta$ lens*.

Another common requirement is for *telecentricity* in image space, i.e. the central, or chief rays of the beam are parallel with the lens axis for all scan angles. For a thin lens it can be seen from figure C2.7.11 that this is achieved when the entrance pupil of the lens is positioned at the back focus of the lens. Telecentricity ensures that the focused spot is not elongated into an ellipse by the angle of incidence at the image plane. The drawback is that at least the final lens element must be longer than the scan line. In practice, true telecentricity is often not required, and a reasonable divergence of chief rays from the scan lens can be tolerated, greatly reducing the size of lens elements. The fractional increase of spot dimension in the y-direction is

$$\frac{\Delta s}{s} = \frac{1}{\cos(\theta)} - 1 \tag{C2.7.17}$$

so, for example, an angle of 15° increases the spot dimension by only 3.5%. In practice, this would be acceptable in many systems.

With modern lens design software, it is quite possible to design lenses that are simultaneously corrected for field flatness, $f\text{–}\theta$ distortion and telecentricity. However, excellent optical correction is also usually required, to ensure uniform spot size over the scan line. The fewer the constraints placed on the design, the better the achievable lens resolution performance. Although it is sometimes possible to

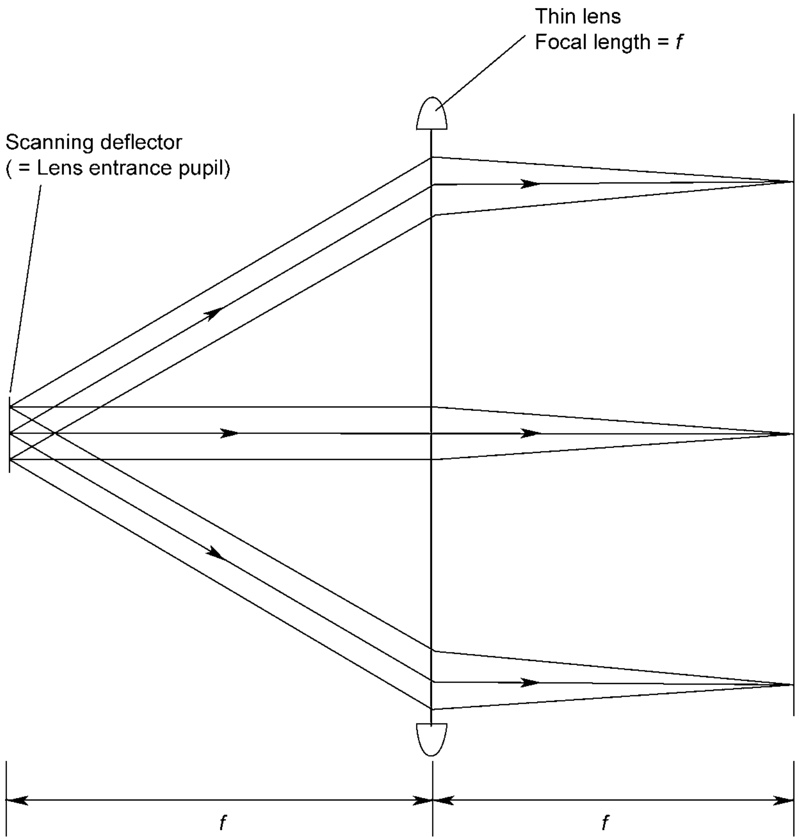


Figure C2.7.11. Telecentric scan lens.

design around an existing lens design, the system-specific requirements for wavelength, scan line length and resolution most often dictate a need for a custom lens.

Conventional, spherical-surfaced glass lenses typically comprise three or more lens elements. Although these can be of a single glass type, better correction is obtained, even for a monochromatic requirement, by using more than one glass. Using a combination of crown and flint glasses, colour correction can be achieved for two, or even three, wavelengths. Usually, diffraction-limited performance is required; this becomes more difficult to achieve as the spot size decreases and the scan angle increases. The number of resolved spots per scan line is a good measure of the lens quality, with anything over 25 000 spots/scan line being difficult to achieve.

For systems that are manufactured in large numbers, one- or two-element injection-moulded polymer lenses are usually used. As the mould tool determines the lens shape, aspheric surfaces are economic, and the lens can therefore achieve a good optical performance with fewer surfaces than can be achieved with spherical surfaces. By using very complex, asymmetric surfaces, it is possible to correct even the effects of pupil wander—the small movement of the beam relative to the lens axis during polygon rotation [23].

Because of the limited choice and range of materials, colour correction is not possible with plastic lenses, and the number of spots per scan line is usually less than 10 000. Thermal stability is often an issue, even within a normal office environment, because of the high temperature coefficients of expansion and refractive index of plastics. It is possible to passively compensate for temperature variations by carefully designing hybrid glass/plastic optics [24].

C2.7.5.4 Deflector error correction

A number of design techniques have been developed to reduce, or eliminate, cross-scan deflection errors caused by manufacturing and alignment tolerances of prismatic polygons. These techniques allow relaxation of what would otherwise be prohibitively tight tolerances.

Active deflection error correction uses a small-angle, high-speed deflector in the optical path just before the main scanning polygon. This controls the input beam angle in the cross-scan direction, to compensate for cross-scan angle errors in the polygon. In an open-loop control system, the fixed pyramidal errors of the polygon mirror are programmed in a calibration procedure. In a closed-loop control system, a separate optical measurement system is used to measure the cross-scan deflection angle, and a servo control system dynamically adjusts the input angle to correct any deviation. In the latter case, non-repeatable errors such as bearing wobble can be corrected in addition to polygon fabrication errors.

Passive deflector correction can be achieved at a much lower cost using non-rotationally symmetric optical surfaces in the flat-field lens. The principle of the correction system is to use a cylindrical lens to focus the input beam on the polygon facet only in the cross-scan plane, and then to use cylindrical or toroidal elements within the flat-field lens to make the facet surface conjugate with the image plane in the cross-scan direction (figure C2.7.12). Where the optics consists of conventionally polished glass optical elements, a toroidal lens is added to the usual flat-field lens either before the lens, or close to the image plane. However, where moulded plastic optics is used, the toroidal elements, incorporating aspheric surface forms, may be integrated into the lens design, minimizing the number of optical elements and achieving better cross-scan error correction.

Other correction systems are possible, using multiple reflections to cancel the deflection errors. A pentaprism has the well-known property that the deflection angle is a constant right angle, independent of the angle of incidence, and this is generally true for double-reflection systems. A pentaprism [25], or any of its open mirror equivalents [26], might be used as the deflector element of an internal drum scanner. These internal correcting reflector systems have the disadvantage of requiring significant complexity in mechanical mounting and balancing. This disadvantage does not apply to external correction systems. An external double-reflection element may be used to correct deflection error in a polygon. In practice, this greatly compromises the system design, significantly increasing the polygon size and flat-field lens aperture and reducing scan duty cycle.

C2.7.6 Input scanning systems

In this section, we look at examples of designs of scanning systems used for input of data. These are essentially electronic imaging systems, but they can give far more information than a visual image, providing precise measurement of colour, scattering properties, size or fluorescence at all points within the scanned area.

Scanning systems have been developed for one-dimensional (e.g. supermarket barcode scanners and wire diameter measurement systems), two-dimensional (e.g. desktop document scanners) and three-dimensional (e.g. reverse engineering of aerodynamic automobile models) measurements. Three-dimensional scanners are mainly one of the three types. The first type uses laser radar principles, i.e. a two-dimensional spatial scan combined with high-speed time-of-flight measurement of fast laser pulses.

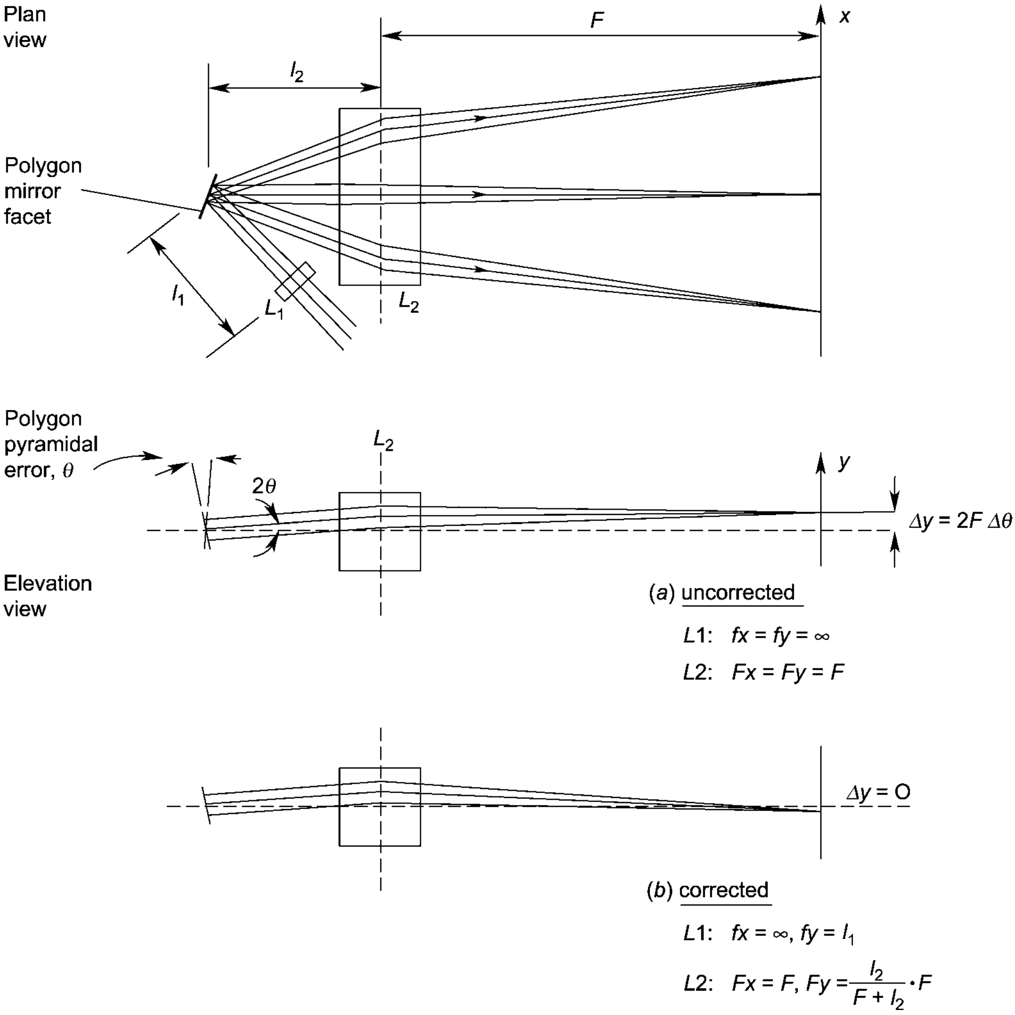

Plan view
F
x
l_2
Polygon mirror facet
l_1
L_1
L_2
Polygon pyramidal error, θ
L_2
2θ
y
$\Delta y = 2F\,\Delta\theta$
Elevation view
(a) uncorrected
L1: $f_x = f_y = \infty$
L2: $F_x = F_y = F$
$\Delta y = 0$
(b) corrected
L1: $f_x = \infty$, $f_y = l_1$
L2: $F_x = F$, $F_y = \frac{l_2}{F + l_2} \cdot F$

Figure C2.7.12. Use of anamorphic lenses for deflection error correction.

The second type uses optical triangulation to measure step changes in the height of a projected laser line, measured by a two-dimensional sensor array at an angle to the line projection plane. Lastly, confocal scanning microscopy [27, 28] uses a small-aperture detector at the focus of a retro-reflected beam to produce a scanned image with very small depth of focus, enabling high-resolution three-dimensional images when the object is stepped through focus on successive scans.

Input systems can be classed as either passive (i.e. using ambient or fixed illumination with a scanning detector) or active (i.e. providing illumination via a scanned beam for a fixed or scanning detector). FLIR thermal imaging systems [29] detect infrared thermal emissions in the 8–12 μm band. Active input systems that use the scanned illumination beam (usually a laser) to define the scanned image resolution are known as flying spot scanners. Examples of applications that use these two classes of scanning are discussed later.

C2.7.6.1 CCD document scanner

CCD scanners are based on linear array CCD detectors, which image the width of the document, with a linear mechanical scan stepping along the length of the document. Scanners are classed as either contact type, where the resolution is essentially that of the detector elements, or reduction type, in which higher resolution is achieved by imaging the document on the detector array with a magnification less than unity. Reduction-type scanners have higher performance, but are more expensive and bulky.

A schematic optical layout of a typical large-format document scanner is shown in figure C2.7.13. Illumination along a line of the document (which could, for example, be a poster, a map or an engineering drawing) is provided by two cylindrical fluorescent lamps, concentrated by cylindrical reflectors. In a camera assembly, a multi-element lens images the line on a tri-linear CCD device. This simultaneously produces three red, green and blue colour-separated images of a line of the document, immediately after the time-delays between the lines are corrected. This provides sufficient data to accurately measure the colour of the document at each point, after calibration.

The spatial resolution is determined by the number of elements in each line of the CCD array, typically 5000, and the lens magnification. The elements are typically 8 μm square, and there is no gap between elements in a line. To obtain higher resolution, two or more camera systems are optically butted together, to increase the number of resolved spots per scan line. This is achieved by slightly overlapping the images of the two CCD sensors at the document, and 'stitching' the electronic images together by software processing.

The CCD array scans the line image data electronically, and the line is scanned along the document by a motorized pair of pinch rollers that transport the document across the stationary optical system. To achieve the necessary precision in this transport system, a sophisticated control system is required to ensure accurate speed and constant tension in the scanned document.

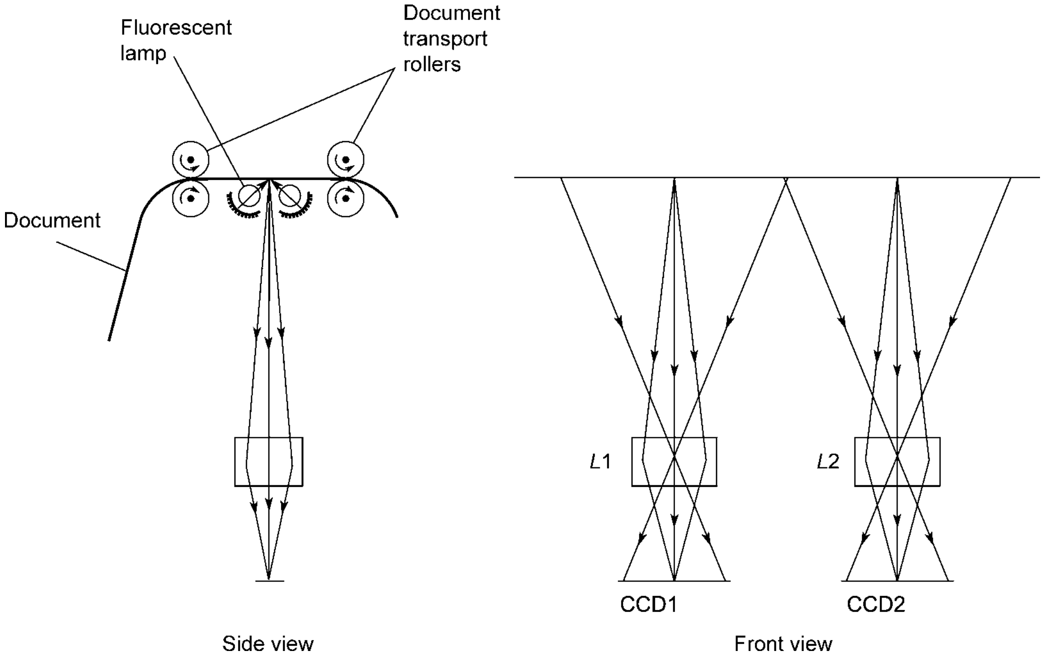


Figure C2.7.13. Large-format document scanner.

C2.7.6.2 Flying spot scanners

The term flying spot scanning was originally applied to CRT scanning, but nowadays it is more often applied to laser scanning. The resolution of the system is defined by the laser spot size at the object, and a non-imaging detector system collects the light that is transmitted, reflected or scattered by the object. Laser barcode scanners work on this principle.

A more complex system, used for on-line inspection of defects in float glass, is shown in figure C2.7.14. In this system [30], a 633 nm HeNe laser beam is scanned by a rotating 12-facet prismatic polygon across the 2 m wide glass sheet during manufacture, as the glass is moved on a conveyor in a direction perpendicular to the scanning direction, producing a continuous raster scan of the sheet. Glass defects are detected and analysed on-line by measurement of the transmission and scatter of the 0.45 mm diameter laser spot.

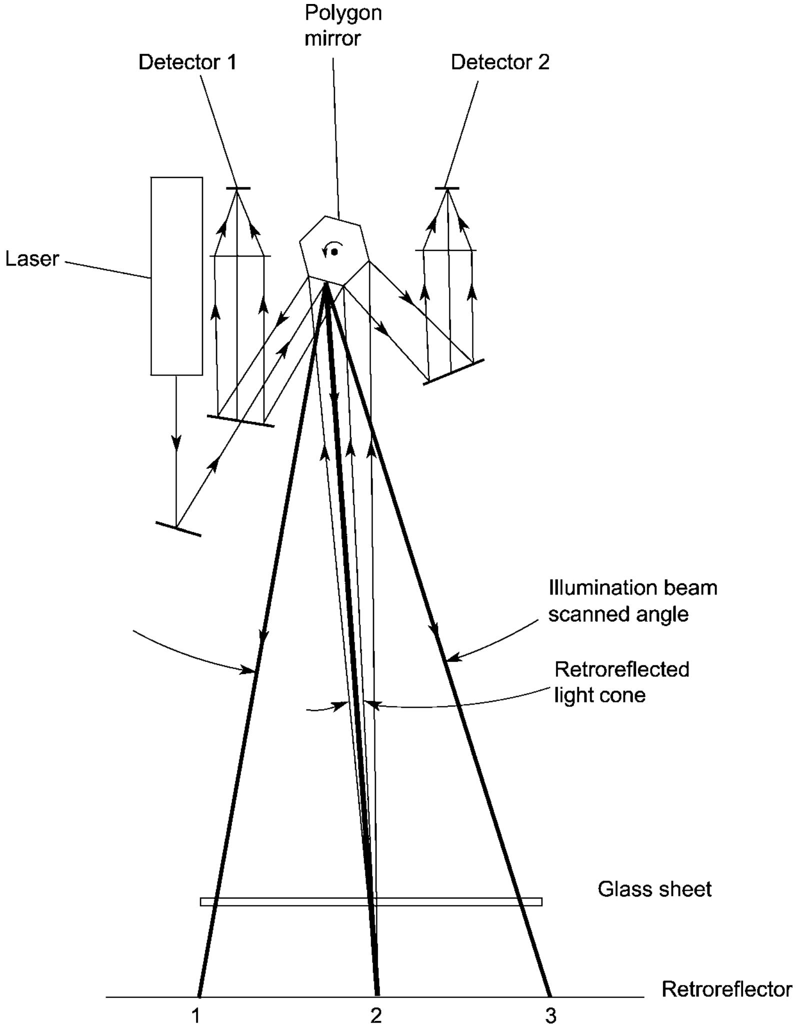

Polygon mirror
Detector 1
Detector 2
Laser
Illumination beam scanned angle
Retroreflected light cone
Glass sheet
Retroreflector
1
2
3

Figure C2.7.14. Glass sheet inspection scanner.

In this system, the detector system is contained within the scanner head, by employing a retro-reflecting screen of microspheres to return the light transmitted through the glass to the scanning system. The reflected light forms stationary images of the scanning spot at two detector positions, returning via two facets of the scanning polygon. One of the detectors is partially masked by a circular obstruction corresponding to the laser spot size, and this measures the scattered light. The other detector simply measures the transmission through the glass.

The polygon facet size is determined by the light collection requirements of the detector system, rather than by the size of the illuminating laser beam. As the retroreflecting sheet produces a cone of about 2° angle, a large polygon with facets of 50 mm × 25 mm is required, which makes the polygon manufacture challenging.

C2.7.7 Output scanning systems

In output scanning, a light source (almost always a laser) is used to write image data to a screen or to a light-sensitive output medium. Systems vary greatly, depending on the size and characteristics of the laser, whose wavelength can be between 350 nm and 10.6 μm, and whose power can be between 1 mW and 1 kW.

Most output systems are two dimensional, imaging to a flat or cylindrical surface. Image size, resolution and speed are the key scanner characteristics. Laser displays take advantage of the eye's persistence of vision to create a two-dimensional image for advertising or entertainment.

The writing process can be based on thermal (e.g. laser marking) or photonic (e.g. photographic) processes. In some processes, the image density is a monotonic function of incident power density, in which case the image is described as continuous tone, or contone. Other processes are sharply thresholded, so that the image density takes one of two (or rarely, more) discrete values depending on the incident power density. In the latter case, the image is made up of a fine array of dots, and is described as halftone.

The power density variations are produced by modulating the continuous-wave laser power as it is scanned across the image surface. Modulation may be controlled by an external acousto-optic or electro-optic modulator, or by controlling the laser current in some cases, particularly in laser diodes.

C2.7.7.1 Desktop laser printer

A typical xerographic laser printer is shown schematically in figure C2.7.15. The configuration may be described as a single-pass external drum, or as a flat-field line scan, with drum rotation providing the slow axis scan. In this example, two adjacent scan lines are imaged from independent low-power 780 nm laser diodes LD1 and LD2, doubling the imaging (drum rotation) speed. The image is produced by a spatial variation of electrostatic charge on the pre-charged photoconductive drum surface, which is discharged by incident light. Toner particles are attracted to the charged areas, and transferred by contact to sheets of paper to produce the finished print.

In the example shown, an eight-facet polygon produces the scan, rotating at 30 000 rpm. A three-element $f-\theta$ lens produces an elliptical spot of 50 μm in the scan direction and 60 μm in the cross-scan direction, for a printer resolution of 600 DPI over a scan length of 312 mm. This allows a printing speed of 20 A3 copies per minute.

Key features of the design are:

- The two collimated laser diode beams are efficiently combined using a polarising beamsplitter, aligned in such a way that the beams overlap at the polygon—which is also the entrance pupil of the $f-\theta$ lens—and are at an angle to each other to separate the beams by exactly one scan line at the image drum.

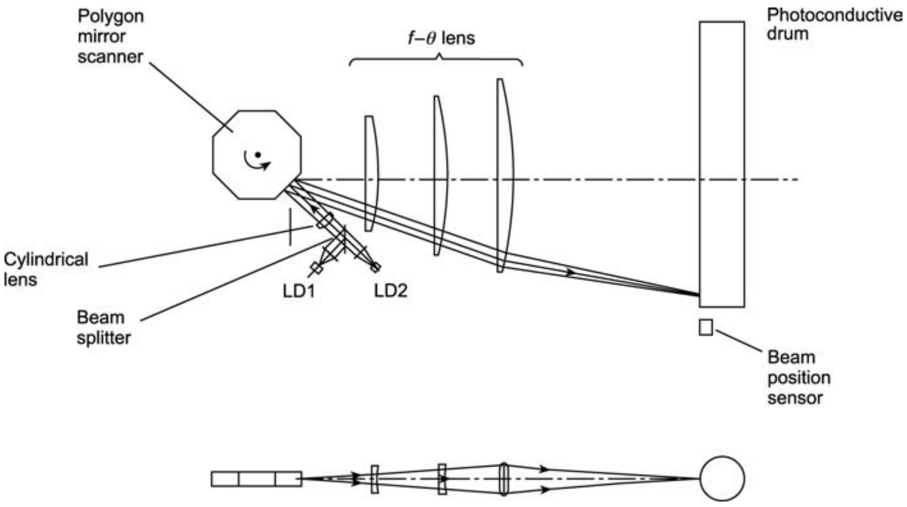


Figure C2.7.15. Laser printer.

- A photodiode just beyond the image area monitors the beam power and position, as part of the control feedback loops for the laser diode and polygon motor drivers.
- Deflector error correction is achieved using an anamorphic lens design, with a cylindrical lens before the polygon and different $f-\theta$ optical power in in-scan and cross-scan planes.
- One of the $f-\theta$ lens elements is glass, with spherical surfaces. The other two elements are injection-moulded plastic with anamorphic, aspheric surfaces. The use of glass enables thermal compensation of the lens performance over a 10–40°C range, which would be a problem with an all-plastic lens.

References

[1] Beiser L 1974 Laser scanning systems *Laser Applications* vol 2, ed M Ross (New York: Academic) pp 53–159
[2] Levy L 1968 *Applied Optics* vol 1, (New York: Wiley)
[3] Marshall G F 1999 Risley prism scan patterns *Proc. SPIE* **3787** 74–87
[4] Boreman G 2001 *Modulation Transfer Function in Optical and Electro-Optical Systems* (Bellingham, WA: SPIE Press)
[5] Montague J 1991 Galvanometric and resonant low-inertia scanners *Optical Scanning*, ed G F Marshall (New York: Dekker)
[6] Brosens P 1976 Scanning accuracy of the moving-iron galvanometer scanner *Opt. Eng.* **15** 95–98
[7] Verboven P E 1988 Distortion correction formulas for pre-objective dual galvanometer laser scanning *Appl. Opt.* **27** 4172–4173
[8] Calquhoun A B, Gordon C S and Shepherd J 1988 Polygon scanners — an integrated design package *Proc. SPIE* **96** 184–195
[9] Beiser L 1991 Design equations for a polygon laser scanner *Proc. SPIE* **1354** 60–66
[10] Sherman R 1991 *Optical Scanning*, ed G F Marshall (New York: Dekker)
[11] Ketabchi M, Tiffanhy B L and Vettese T J 1997 Single-facet scanning subsystem for digital imaging *Proc. SPIE* **3131** 290–299
[12] Ng R K M, Rockett P and Wardle F P 1999 Analysis of the distortion of a high-speed single-facet rotating mirror *Proc. SPIE* **3787** 252–263
[13] Cindrich I 1967 Image scanner by rotation of a hologram *Appl. Opt.* **6** 1531–1534
[14] Kramer C J 1991 Holographic deflectors for graphic arts systems *Optical Scanning*, ed G F Marshall (New York: Dekker)
[15] Kay D B 1984 Optical scanning with wavelength correction *US Patent* 4 428 643
[16] Fowler V J 1964 Electro-optical light beam deflection *Proc. IEEE* **52** 193
[17] Li J, Chen H C, Kawam M J, Lambeth D N, Schlessinger T E and Stancil D D 1996 Electrooptic wafer beam deflector in $LiTaO_3$ *IEEE Photon. Technol. Lett.* **8** 1486–1488
[18] Gottlieb M 1991 Acousto-optic scanners & deflectors *Optical Scanning*, ed G F Marshall (New York: Dekker)
[19] Korpel A 1981 Acousto-optics *Appl. & Opt. Eng. VI*, ed R Kingslake and B J Thompson (New York: Academic) pp 89–141

[20] McManaman P F and Watson C A 1997 Optical beam steering using phased array technology *Proc. SPIE* **3131** 90–98

[21] Stockley J E, Subacius D and Serati S A 1999 Liquid crystal on VLSI silicon optical phased array *Proc. SPIE* **3787** 105–114

[22] Levrentovich O D, Shiyanovskii S V and Voloschenko D 1999 Fast steering cholesteric diffractive devices *Proc. SPIE* **3787** 149–155

[23] Li Y and Katz J 1997 Asymmetric distribution of the scanned field of a rotating reflective polygon *Appl. Opt.* **36** 342–352

[24] Yamaguchi M and Shiraishi T 1999 Development of four-beam laser scanning optical system *Proc. SPIE* **3787** 2–12

[25] Starkweather G K 1984 *US Patent* 4 475 787

[26] Marshall G F, Vettese T J and Carosella J H 1991 Butterfly line scanner *Proc. SPIE* **1454** 37–45

[27] Wilson T (ed) 1990 *Confocal Microscopy* (London: Academic)

[28] Masters B R (ed) 1996 *Selected Papers on Confocal Microscopy*, (Bellingham, WA: SPIE Press)

[29] Lloyd J 1975 *Thermal Imaging Systems* (New York: Plenum)

[30] Holmes J D 1997 Inspection of float glass using a novel retro-reflective laser scanning system *Proc. SPIE* **3131** 180–190

[31] Beiser L 1983 Generalised equations for the resolution of laser scanners *Appl. Opt.* **22** 3149–3151

[32] Aylward R P 1999 Advances and technologies of galvanometer-based optical scanners *Proc. SPIE* **3787** 158–164

[33] Shinada H 1996 *US Patent* 5 502 709

[34] Beiser L 1979 *Holographic Scanning* (New York: Wiley)

C3.1
Optical fibre sensors

John Dakin, Kazuo Hotate, Robert A Lieberman and Michael A Marcus

C3.1.1 Introduction

It is said that Man has five senses—sight, touch, smell, taste and hearing—at his disposal. Clearly, the one he finds most generally valuable is his sight. This is hardly surprising, as it is excellent for remote sensing, gives an effectively instantaneous response, has truly enormous parallel information capacity, and provides far more reliable and quantitative data than any of the others—it is common to say 'you believe the evidence of your own eyes'. Optical sensor technology may still have some way to go to match the same compact design and all-round performance as the eye/brain combination, but the promise is clearly there! As an aside, it is worth noting that we, as humans, are not able to remotely sense electrical or magnetic fields well (although migrational birds are believed to use magnetic sensors) so, if one believes that nature often chooses the best methods, the technology of optical sensors may perhaps evolve to overtake that of electrical ones in time. Modern camera and CD player technology has shown that even highly complex optoelectronic systems can be manufactured cheaply in volume, so cost should not present a major barrier for large-scale application of standard devices.

This chapter reviews the many ways in which optical fibres may be used, in conjunction with optoelectronic instrumentation, to sense physical or chemical parameters. This is done using either intrinsic sensors, where the fibre itself acts as the sensing element, or extrinsic sensors, where the fibre acts in its normal communication role, to carry light to and from an external optical sensor. The fibre can enable remote and rapid sensing to be performed, by sensing at specific points, or over specific cable paths. This complements the capability of optical sensors such as cameras which have valuable imaging capability, as fibre cables do not need a direct 'line-of-sight' to operate and can measure in inaccessible places. Because there are now very many varieties of optical fibre sensor, it has been necessary to concentrate on a few important concepts and only describe a few practical sensor types in more detail. A far more detailed account of fibre sensors is given in the textbooks mentioned in bibliography references 1 and 2.

Optical fibres are widely used in long distance trunk telecommunications systems (see section C1) and their use in shorter distance applications is growing rapidly. For communications systems, it is desirable to utilize a cable having transmission properties that are insensitive to environmental changes. Intrinsic fibre sensors, in contrast, rely on deliberately configuring the optical system to be sensitive to external influences on the fibre or cable. This involves the arrangement of the fibre cable, its associated light source, the optical receiver and the chosen signal processing scheme to maximize (and detect) a change in transmission properties that is characteristic of the parameter to be sensed. This may be carried out either by using especially sensitive cables, or by interrogating conventional fibre cables in a manner that highlights small changes in transmission. The resulting changes in transmission can then be used to measure features of the external environment.

An extrinsic fibre sensor usually uses a conventional optical fibre (preferably sheathed within an environmentally insensitive cable) which is only employed as a convenient light-guiding medium to transport light to and from more conventional optical sensors at the end of, or at specific points on, the cable. This sensor, in response to an external physical parameter, modifies the light coupled back from the input fibre into the return fibre and guided back to the detector system. Such a sensor may still be undesirably affected by environmentally induced changes in cable properties, unless particular care is taken with its design. Avoiding unwanted cable-induced errors is a very important practical consideration, in both intrinsic and extrinsic sensors. In a good sensor system design, care will be taken to avoid significant sensitivity to undesirable losses in the optical network, such as could occur, for example, due to fibre bending, or due to connector variations.

The research area of fibre optic sensors is a very fruitful one for original thought, as there is not only a very large number of physical and chemical parameters which may be sensed, but also a multitude of ways in which the parameter may be arranged to modulate the optical transmission. However, in order to make a successful sensor, the normal principles of good instrument design must be adhered to (i.e. low crosstalk to undesirable parameters, good signal-to-noise ratio (SNR) and repeatable and reliable operation) and the sensor must be designed to be cost effective for the intended application. Many of the successful commercial applications of such sensors are in niche market areas, which rely on the inherent advantages of optical sensors over more conventional electrically based sensors. The primary technical advantages include the following:

- Intrinsic freedom from electromagnetic interference (EMI), lightning strike, etc.
- Intrinsic safety in hazardous (explosive vapour) environments (provided optical signal power is low, which is almost invariably the case).
- High electrical isolation, enabling their use in medical applications and for data collection from points at high voltage. The fibre also gives freedom from problems of electrical short circuit and open circuit.
- Excellent resistance to chemical corrosion (can be used in highly corrosive environments, e.g. saltwater, acid, alkali, etc).
- Passive operation; no electronics circuitry or electrical power is required at the remote sensing point.
- May be used in high temperature areas, where electronic systems would not survive.
- Optical fibres are smaller, lighter and cheaper than electrical cables.
- May be used for distributed sensors of extreme length, due to the low losses of ($0.2\,\mathrm{dB\,km^{-1}}$) achievable in optical fibres.
- Where optically based data transmission is already envisaged, sensor information can be carried in the same optical fibres.
- Optical sensing can provide a very rapid response in many applications, as changes in optical transmission can be detected almost instantly.

When the use of optical signal cables becomes more common for short distance routes, it is likely that optical transducers will start to find more general applications, even in areas where the above intrinsic advantages are not a major consideration. The ability to sense, communicate and multiplex signals within an optical network is an important attribute, which will lead to their greater application.

We shall start by considering the different types of sensor, and the physical means by which each can operate.

C3.1.1.1 Summary of intrinsic types of sensors, where the optical fibre is used directly as the sensing element

The *intrinsic* optical fibre sensor takes advantage of measurable changes in transmission characteristics of the fibre itself. The principal parameters of interest for sensing are:

- Light generation in the fibre due to physical interactions (e.g. scintillation or Cerenkov radiation).
- The propagation time of light in the fibre (proportional to the length and inversely proportional to the velocity of light). This can be measured as a temporal delay or as a phase change in the light.
- The optical power transmitted by the fibre (either the total power or the spectral variations in transmission).
- The distribution of optical power between the various modes of propagation. (This can be measured from either the near- or far-field waveguide mode patterns at the end of the fibre.)
- The state of polarization of the transmitted energy through the fibre (or backscattered energy from the fibre).
- The light scattered from within the fibre core material. This can include elastic scattering (same wavelength as incident light) or involve Raman scattering, Brillouin scattering, or other nonlinear interaction, such as optical gain by Raman or Brillouin processes.

C3.1.1.2 Summary of extrinsic types of sensors, where the optical fibre is used merely as light guiding medium, to address more conventional optical sensors

The main methods of operation of *extrinsic* fibre sensors are outlined below:

- Light due to a chemical or physical interaction is collected by the fibre, e.g. radiation due to high temperature pyrometry, optical scintillation of a semiconductor, chemiluminescence).
- The optical power transmitted by an external optical modulator coupled to the fibre (either the total power or the spectral variations in transmission) is monitored. Examples include mechanical optical shutters, electro-optic switches or modulators, variable optical filters (such as Fabry–Perot, grating monochromator, temperature-dependent semiconductor).
- The optical power reflected by an external optical modulator coupled to the fibre (either the total power or the spectral variations in transmission) is monitored. Examples as above.
- The optical power scattered elastically (no wavelength change) by an external medium coupled to the fibre is monitored. This effectively measures turbidity of the medium, in reflective mode.
- The optical power scattered inelastically (scattered-wavelength changes) by an external medium coupled to the fibre is measured. This includes processes such as Raman scattering, fluorescence and photon correlation spectroscopy.
- The state of polarization of the energy transmitted by an external optical modulator coupled to the fibre (or that reflected back into the fibre) is monitored.

C3.1.1.3 Evanescent field sensors

Evanescent field sensors exhibit some of the characteristics of *intrinsic* sensors and *extrinsic* types, but do not exactly fit into either category. As with *intrinsic* sensors, the light is guided by the fibre (or by an optical waveguide attached to the fibre) in the sensing region, but, in the case of *evanescent field* sensors, a portion of the optical energy travels outside the physical limits of the waveguide material. In the lateral direction, the field decays away rapidly as distance from the waveguide increases, a behaviour known as *evanescent field* decay. Because there is light energy outside the guide material, it is possible for this light to interact with (e.g. be absorbed, scattered, or excite fluorescence in) the surrounding material. Then, for example, the *evanescent field* sensor can detect the optical properties of a fluid in which it is immersed or solid material in which it is embedded, simply by measuring the light transmitted through the fibre. Unfortunately, because the evanescent field region is very thin (typically less than 1 μm thick), this type of sensor is very prone to surface damage or contamination. Also, the evanescent field penetration depends strongly on the refractive index of the surrounding material (and hence on temperature) and on whether or not the guide is bent, so is very sensitive to the operational environment.

Instead of immersing the evanescent field sensor directly into a solution to be monitored, it can be coated with an active layer, such as a polymer or a semi-permeable, glass-like, sol–gel coating. These coatings can be made sensitive to desired chemicals by incorporating (immobilizing) an indicator chemical into them. Provided the layer is semi-permeable, chemicals can diffuse in from the surroundings and change the optical transmission of the indicator material. Clearly the presence of this layer reduces the problems of contamination of the critical evanescent field region, as this now lies inside the solid material.

A further variation is to coat the light guide with a thin metal layer, which can enhance the field in the evanescent region by a mechanism called plasmon resonance, which, as the name implies, involves excitation of the ‘electron gas’ present within all metallic conductors. This plasmon resonance mechanism can give greatly improved sensitivity to any absorption in the evanescent field region, and still allows indicator layers to be coated on as before.

Because of the ease of contamination and the environmental sensitivity of evanescent sensors, they are usually better suited to qualitative testing for chemicals, rather than a means of performing quantitative chemical or spectral analysis of fluids in which they are immersed. Nonetheless, they can be a very sensitive means of detecting trace quantities of chemicals, provided they are designed and used with care.

The following types of *evanescent field sensor* (with bracketed notes to indicate where more details can be obtained from bibliography 1) are commonly used:

- Unclad fibre, with evanescent field extending directly into the gas or liquid to be sensed (bibliography 1, vol 1, p 607).
- Fibre with a physically sensitive (or chemically-reactive) cladding layer, with the evanescent field extending into cladding. The chemical must diffuse into, and physically modify, or react with, the layer (bibliography 1, vol 4, pp 352–354).
- Polished half-coupler sensor. Here the fibre is set in a slit in a glass block, with part of the cladding polished away to expose the external medium (or a reactive polymer coating layer) to the evanescent field region (bibliography 1, vol 1, pp 213–215).
- Integrated optics sensors, with natural propagation of evanescent field above surface of chip. Again there is exposure of the external medium (or reactive polymer layer) to the evanescent field region (bibliography 1, vol 1, chapter 9).

- Plasmon resonance. Here, a thin metallic layer enhances the optical coupling of the light in the fibre to the evanescent field region in the external medium (or reactive polymer layer). This enhancement can be used with most of the above configurations (bibliography 1, vol 1, pp 203–206).

Please note that chapter 3, by G Stewart, in vol 3 of bibliography 1, gives a detailed overview of evanescent field devices.

Below, we shall give more detailed descriptions and case studies of optical sensors. The remainder of the chapter is split into four sections, each written by a different co-author. Section C3.1.2 (John Dakin) covers the basics of intensity-based sensors and section C3.1.3 (Kazuo Hotate) treats interferometric types. Section C3.1.4 (John Dakin) covers how sensors may be multiplexed and section C3.1.5 (John Dakin, Bob Lieberman and Kazuo Hotate) shows how sensors that operate on a fully distributed basis can be made. The final section (Mike Marcus) deals with a detailed case study of a type of interferometric sensor that has many industrial uses. Because the chapter is composed of contributions from several authors, there are small differences in style of writing and diagrams reflecting their different contributions. It should be emphasized again that not all types of sensor can be covered in such a short chapter, and there has had to be much careful selection of which types to include and which case studies to present in more detail. The reader requiring more detail is again referred to the bibliography, to the 'Optical fibre sensors' and the 'Europt(r)ode' international series of conferences and also to the many SPIE-organized (www.SPIE.org) conferences covering this area.

C3.1.2 Intensity-based optical fibre sensors

C3.1.2.1 Physical sensors

Simple optical intensity sensors

We shall first describe the very simplest form of intensity-based sensors, where only the transmitted or reflected light level is measured. We shall start with intrinsic sensors, then discuss extrinsic types.

Light generation in fibre itself

A few mechanisms allow generation of light in a fibre, as a result of a direct interaction with the physical field to be monitored. There are two principal types here. One is a fibre sensor for detection of pulses of ionizing radiation, having extreme peak intensity, of the type that can occur due to thermo-nuclear events (see figure C3.1.1).

If a relativistic charged particle (one with a velocity in free space close to the velocity of light in vacuum, c) passes through a medium such as glass, of refractive index, n, and if its entrance velocity initially exceeds the velocity of light c/n, in that new medium, then it will lose energy rapidly in the form of electromagnetic radiation. The energy loss occurs due to the Cerenkov effect [1, 2] and is emitted mainly in the form of blue and UV light, but also with some energy extending to longer wavelengths. If an intense pulse of such radiation strikes a fibre, the temporal variations in the pulse intensity can be monitored, by detecting the light at the far end of the fibre.

A sensor that is more sensitive to lower levels of radiation is one that uses fibre fluorescence [3] resulting from the radiation, either fluorescence intrinsic to the silica fibre, or that generated more efficiently by doping the fibre with fluorescent material. Arrays of polymer fibres, doped with fluorescent organic dyes, have been used as long linear scintillators to track the passage of ionizing radiation in particle accelerator systems, with detectors at the end of each fibre to collect the light.

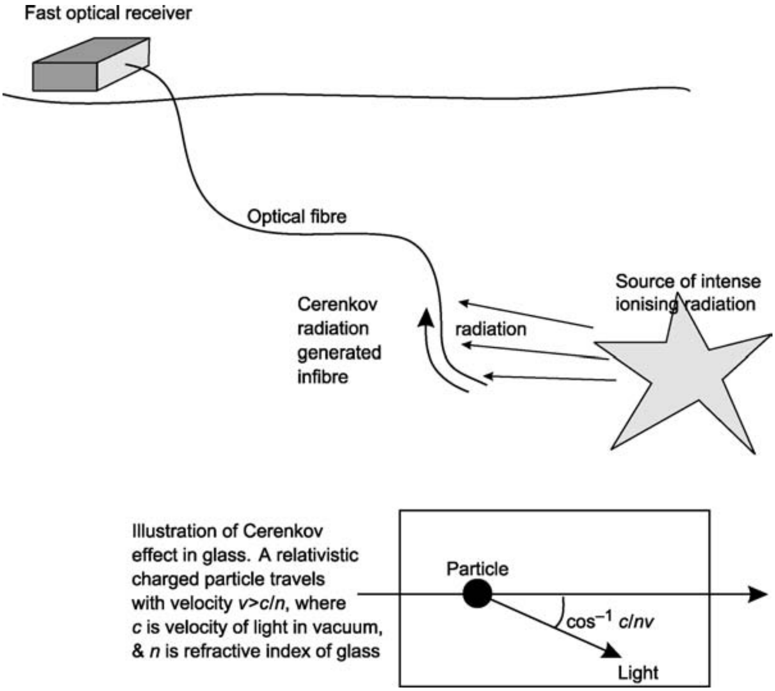

Fast optical receiver
Optical fibre
Cerenkov radiation generated infibre
radiation
Source of intense ionising radiation
Illustration of Cerenkov effect in glass. A relativistic charged particle travels with velocity $v>c/n$, where c is velocity of light in vacuum, & n is refractive index of glass
Particle
$\cos^{-1} c/nv$
Light

Figure C3.1.1. An optical sensor for detection of intense ionizing radiation (relativistic charged particles) using the Cerenkov effect.

Microbend sensors

Microbend sensors [4, 5, 148] are intrinsic sensors, which take advantage of the loss in optical fibres when they are bent. The word microbend is used because gradual bends, i.e. bends of large radius greater than 50 mm, cause very little loss in a fibre, or they would not find the extensive uses that they have in communications, whereas microbends (or sharp 'kinks', of small bending radius of a few mm or less) can cause high losses, even if present over very short lengths.

The losses in the fibre are due to conversion of energy in guided fibre-core modes to cladding modes, which are then usually lost by absorption or scattering in the fibre sheathing material. In the case of monomode fibres, the mechanism is relatively simple, as there is only one guided mode (strictly two, if the duality of polarization modes is taken into account) to be coupled out, and the extent of coupling, and hence the loss, can be uniquely determined by the degree of bending.

In the case of the multimode fibres, more commonly used with such sensors, the situation is much more complex. Firstly, bending causes mode conversion, in particular coupling of lower order modes (associated with rays in the core which travel at a small angle to the fibre axis) to higher order ones (associated with rays travelling at larger angles to the axis) and eventually to cladding modes. This 'chain' of events, coupling energy from low order modes, via higher order ones, to radiation ones, depends in a complex way on not only how the fibre is bent, but also on which modes were initially present in the fibre in the section immediately before the bend. Because of the latter aspects, it is extremely difficult to derive quantitative information from multimode microbend sensors, particularly if several are cascaded along the length of a single fibre, as the bend condition of each one will affect the response of subsequent ones.

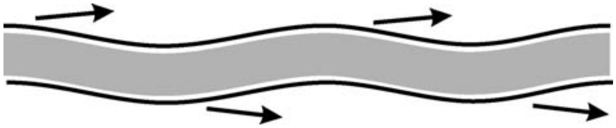
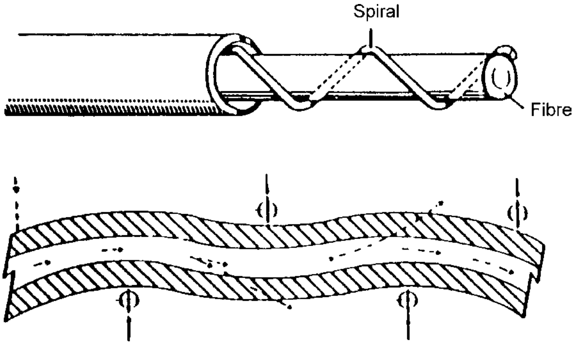


Figure C3.1.2. Schematic showing how light is lost at periodic bends in a multimode fibre.

Multimode microbend sensors have a particularly strong response if they cause the fibre to be periodically bent (figure C3.1.2), with a particular spatial period which corresponds to the 'zig-zag' period associated with the highest order rays that the fibre is capable of guiding [6]. These highest order rays are those striking the core cladding interface at the normal critical angle associated with Snell's law of total internal reflection. Thus, a strong response can be achieved by pressing the fibre between corrugated plates, with offset corrugations (or plates covered with parallel metal pins) which will periodically deform the fibre in a wave-like manner, with the appropriate spatial period (usually around 1–2 mm, for typical multimode fibres) needed to ensure strong mode coupling. Under these conditions, very high losses can be induced in the fibre, even with a transverse displacement of only a few microns amplitude. However, the response is highly nonlinear and care must be taken to avoid the possibility of exceeding the long-term mechanical bend limitations of the fibre. Clearly with soft polymer coatings, such as acrylate, slow mechanical creep will be a problem, leading to variations of sensor response with time. Attempts have been made to improve the mechanical stability and reproducibility, particularly at higher temperatures, by using metal-coated fibres.

Probably the most practically useful, albeit rather qualitative, microbend sensor is the distributed cable version, first devised [7, 8] by Dr Alan Harmer at Battelle labs, Geneva. This consists of an optical fibre which has a thin polymer fibre thread helically wound around it, before the combination is sheathed within an outer polymer tube (see figure C3.1.3). The spatial winding period of the helically wound fibre is designed to correspond with the mode-coupling length. When the outer tube is compressed from opposite sides, the fibre is deformed to cause high loss. This assembly would clearly also suffer from slow mechanical creep, but it finds most use as a cable to sense, essentially in an on/off manner, the presence of lateral force or weight on the cable. Applications are detection of the pressure due to a human foot (e.g. for safety mats near machinery, or for intruder detection on fences or



Principle of the pressure-sensing cable, using microbend sensor. (Courtesy Herga Electric, U.K.)

Figure C3.1.3. The pressure-sensitive cable of Herga Electric, first devised by Harmer. The spirally wound filament causes periodic bending of the inner glass fibre, when the outer cable is deformed by lateral pressure.

perimeters) or as a safety device to detect, and hopefully prevent, trapping of a hand or head, in a sliding door or window (e.g. lift door, automobile window, sun-roof, etc). A variety of such cable and cable-in-mat sensors have been sold commercially by Herga Electric, UK [148].

A more recent development in microbending sensor technology [9, 10] has been to produce a sensor to detect the presence of water around a fibre cable, for example, for the detection of floodwater in cable ducts. This is important, as water can freeze and break fibres, or can cause hydrogen generation by corrosion of metals. The sensor cable design (figure C3.1.4) has a hard inner cylindrical core, coated with a hydrogel layer (so this hydrogel layer has a hollow cylindrical geometry). A light guiding fibre is held against the outer cylindrical surface of this hydrogel layer using a helical plastic fibre to keep it in place. If the hydrogel layer gets wet, it swells dramatically, forcing the optical fibre outwards, except where held in place by the helical fibre. This periodically deforms the fibre to make another form of microbending sensor, in this case one sensitive to water. The diagrams in figures C3.1.2–C3.1.6 illustrate the microbend sensor, plus a number of other loss-based fibre sensors.

Other intrinsic fibre-loss sensors

There are several other types of intrinsic sensor that are based on attenuation in optical fibres.

An attractive sensor [11] is the radiation dosimeter (figure C3.1.5), in which the attenuation of the fibre is increased by ionizing radiation. Silica fibres are suitable for high levels of radiation, as they need large doses to suffer substantial change, whereas lead-glass fibres have a significant response at lower levels. In all cases, however, the losses exhibit a time (and hence a dose rate) dependence and the losses also reduce after thermal annealing. This annealing can be significant even at room temperature.

Another commercially useful sensor [12] for detecting leaks of cryogenic fluids (e.g. liquefied natural gas or petroleum gas) was devised (figure C3.1.6) using a type of commercially available polymer-clad silica (PCS) fibre. The type of fibre used consisted of a pure silica core with a lower

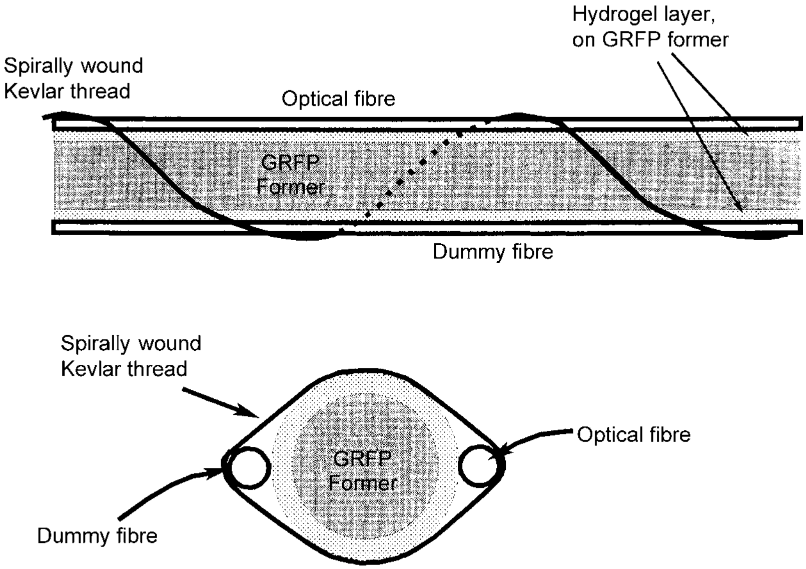


Figure C3.1.4. Illustration of a cable designed for sensing presence of water, which also responds to areas of high humidity [10]. The hydrogel layer absorbs water and swells, causing the glass fibre to be pushed outwards. The outer spiral filament constrains it at regular spatial intervals, causing it to suffer periodic microbending, and hence loss.

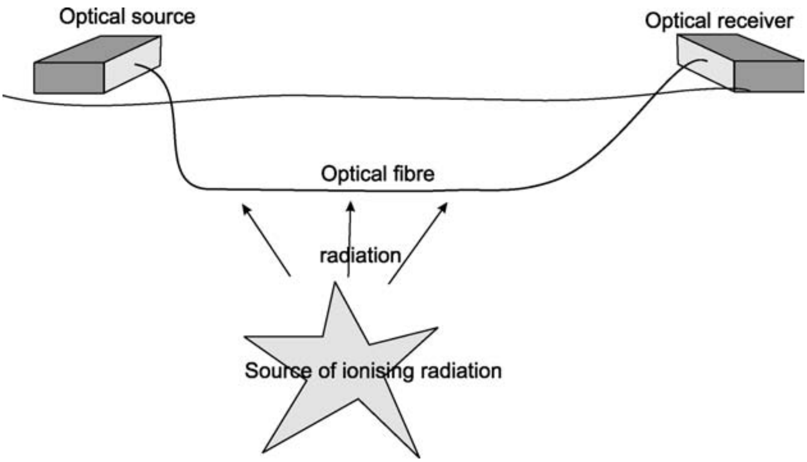


Figure C3.1.5. The optical fibre dosimeter. Ionizing radiation causes increased attenuation in fibres, particularly lead-glass ones, causing the detected signal to reduce.

refractive index optical cladding, composed of silicone polymer material. When cooled to a temperature well below 0°C, the polymer shrinks more than the silica, due to its higher index of thermal expansion, and its refractive index rises to exceed that of the silica core. This causes the attenuation of the fibre cable to increase dramatically, so if light is launched at one end, there is a reduction in the detected light intensity at the output end, thereby indicating a leak of cold fluid at some point along the cable.

Propagation-delay or time-delay sensors

Propagation-delay sensors (figure C3.1.7) allow the monitoring of physical or chemical effects by examining the time light takes to travel through or return from the sensing element.

This element is usually the fibre itself. Its length can change if it is mechanically strained (delay changes arise due to change in physical length or due to refractive index changes arising from the strain) or if it changes temperature. If the fibre is very long, it is possible to examine small changes in the arrival time of a short optical pulse that has travelled through it. However, because of the extremely high velocity of light, it is difficult to perform such a measurement with high precision. An alternative way is

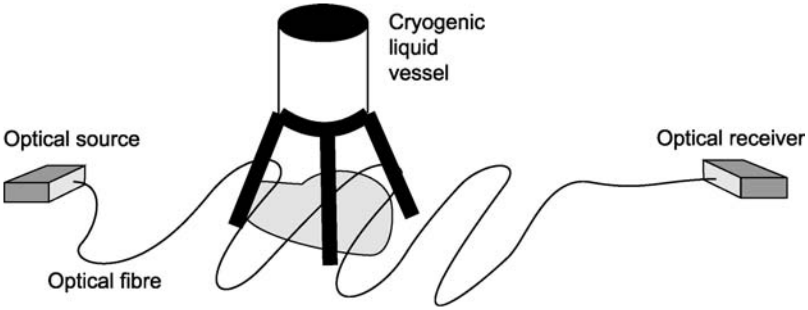


Figure C3.1.6. The cryogenic leak detector of Pinchbeck and Kitchen. Silicone-PCS fibres suffer attenuation when strongly cooled, because the cladding refractive index increases to equal or exceed that of the fibre core, preventing light guidance.

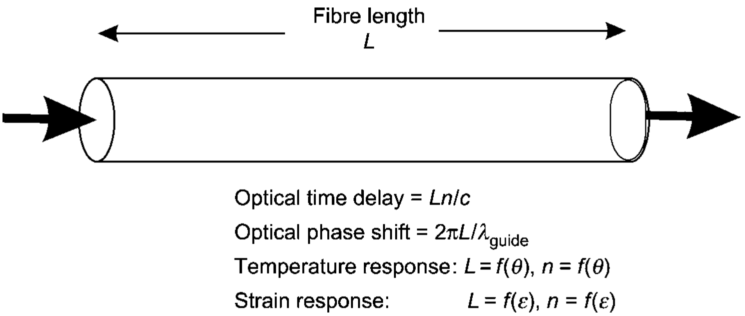

Fibre length
L
Optical time delay = Ln/c
Optical phase shift = $2\pi L/\lambda_{\text{guide}}$
Temperature response: $L = f(\theta)$, $n = f(\theta)$
Strain response: $L = f(\varepsilon)$, $n = f(\varepsilon)$

Figure C3.1.7. The concept of the propagation delay sensor. The delay is a function of both temperature and strain. In intensity-based sensors, the time delay has to be measured by giving the light some form of intensity modulation. (Interferometric detection of delay will be discussed later.)

to modulate the light source at a very high frequency (typically between 500 MHz and a few GHz) and observe phase changes in the intensity modulation envelope caused by delay changes in the optical carrier signal. (Of course, by far the most precise method of monitoring is to launch a continuous lightwave through a length of fibre and examine the changes in optical phase due to the external influence, but this method comes into the area of interferometric sensors, which will be covered in detail in section C3.1.3 of this chapter.)

Extrinsic intensity-based sensors

Light-collecting sensors

An attractive type of extrinsic intensity-based sensor is one in which the fibre merely collects light from an external source and guides it to a detector. The simplest example of this is a fibre used to check that a light bulb (e.g. a vehicle headlight) is operational. Fibres have also been used to monitor the temperature of hot bodies [13] such as turbine blades in jet engines, by collecting Planck hot-body infrared radiation and guiding it to a detector (remote pyrometry). Clearly, without care, there is a risk of surface contamination of the fibre. One alternative to solve this problem [14] is to deliberately cover the fibre end with an opaque metal coating or sealed tube, and collect the energy emitted from the inside surface of the layer or tube, which produces an excellent approximation to a perfect black body (figure C3.1.8).

Later generations of such sensors are now used in commercial instruments. With such fibre pyrometers, the simplest approach is to measure total intensity. The signal conveniently varies with temperature, but even more dramatically than would be expected from the normal 4th power law for the total radiated power, as the fibre detector combination usually only allows transmission and detection of energy from the shorter wavelength, higher energy, 'tail' of the Planck radiation curve, so the measurement accuracy is reasonable even with this simple approach. However, the accuracy can be improved [15] by monitoring at two or more wavelengths, to determine the effective colour temperature of the source. This then reduces the power-referencing errors that can occur when simply measuring intensity on a single detector.

External modulator sensors

Many forms of external modulator can cause variation of coupling between outgoing and return optical fibres. The simplest is just a mechanical shutter between the two fibres, which moves to shut off light. Such simple shutters can act as safety interlocks, e.g. for fire doors, or hazardous areas, where only

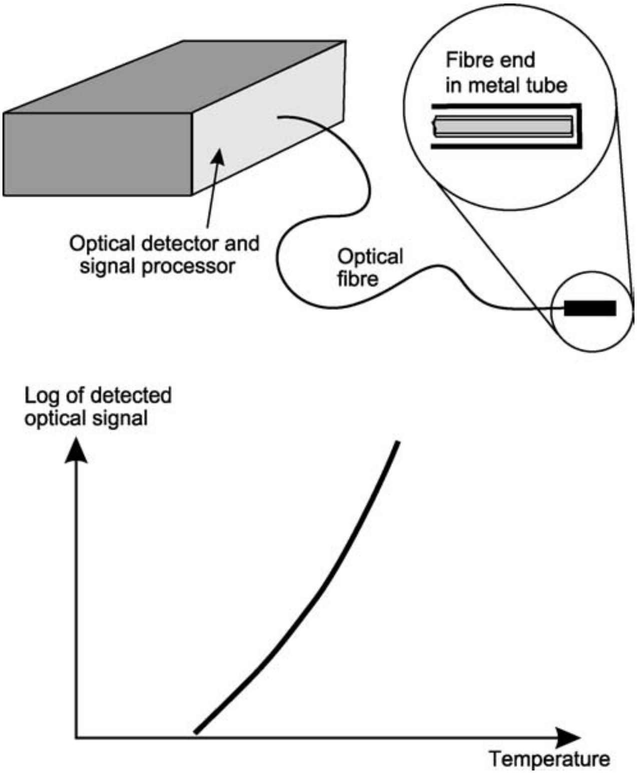


Figure C3.1.8. Fibre pyrometer of Dakin and Kahn, with opaque coating or closed tube over fibre end tip. Such a long coating or tube acts as an excellent black-body source, and is the way commonly used to approximate to a perfect black body. Lower curve shows how the detected signal rises dramatically with temperature.

the safe condition allows light to be transmitted. Such sensors can be connected in series, such that light only passes if all are in a safe condition, making a logical ‘AND’ condition to ensure all are safe. Input light can be modulated to ensure that it is not simply ambient light being collected by breaks in the system. In contrast, electrical switches can always short circuit or open circuit, particularly in corrosive chemical environments, with conductive fluids, or in seawater environments.

A liquid level sensor devised by Pitt [16], involved connecting a small right-angled microprism to the end of a fibre, or of a pair of parallel fibres, such that light was reflected back by total internal reflection when the tip was in air or gas (figure C3.1.9). When wetted by the liquid, the retro-reflection was greatly reduced (as only weak Fresnel refection now occurs) indicating the liquid level has reached the position where the prism lies.

An alternative, in transmissive mode, is a liquid-level sensor [17], made using fibres connected via a small gap, that is filled when the liquid reaches it, and monitoring the transmission at a wavelength where this liquid absorbs strongly (e.g. at 980 nm or 1.4 μm to monitor water). As the transmission change due to absorption can be very strong, particularly for water at 1.4 μm, and particularly because a second (unabsorbed) wavelength of light can used as an intensity reference signal, this sensor is potentially more ‘fail-safe’ in nature than the prism type.

A useful high-resolution proximity/distance detector (figure C3.1.9) can be made simply from two or more parallel fibres, with their end faces terminated in a common reference plane [18, 19]. One or more incident fibre is excited with light from a source, and the return fibre(s) collects the scattered

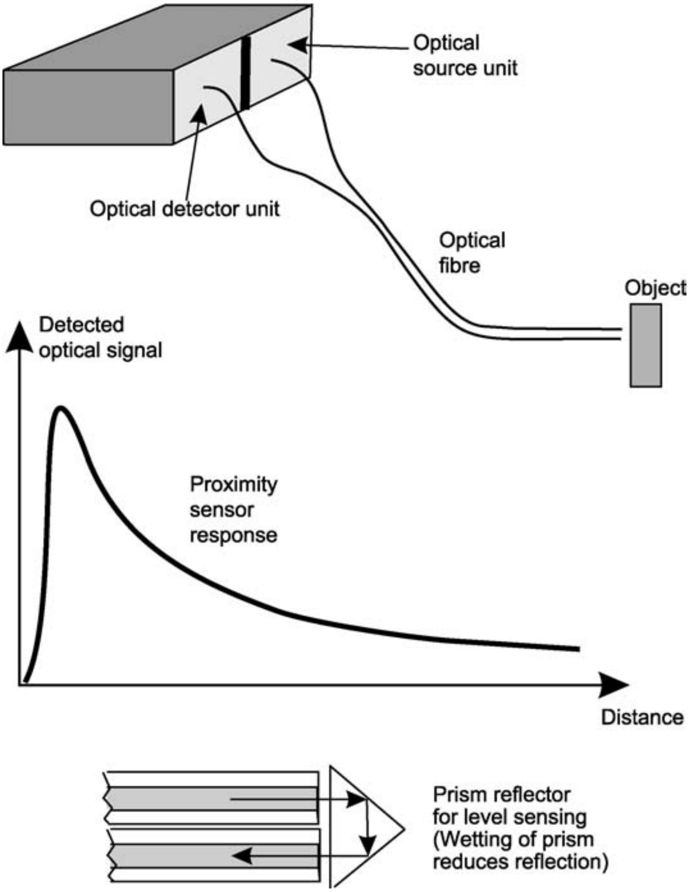

Optical source unit
Optical detector unit
Optical fibre
Object
Detected optical signal
Proximity sensor response
Distance
Prism reflector for level sensing (Wetting of prism reduces reflection)

Figure C3.1.9. The optical fibre proximity detector and its modification to produce a liquid level sensor. If care is taken to avoid the two-way ambiguity in the detected signal from a surface, it can also be used to obtain approximate distance measurement, but errors clearly arise if the reflective properties of the surface vary. Adding a prism to the fibre ends (lower picture) produces a liquid level sensor. In this latter case, immersion prevents total internal reflection and hence greatly reduces retro-reflection by the prism.

or reflected light. Light exits the incoming fibre(s) in an illumination cone, of divergence angle determined by the fibre numerical aperture, and the receiving fibre can only receive light that is reflected or scattered back from a point lying within a similar receiving cone. If a reflective point or surface is too close, no light is received by the return fibre, as these cones do not intersect. As distance to the reflective point increases, there is firstly a very steep increase in signal as the cones start to overlap, then a fairly rapid decrease, this time because return signals reduce due to the much smaller signal that can be received by a tiny fibre core located a long way away from the reflective point. A simple method is to use a split (bifurcated) bundle of many optical fibres, where light is launched through one set and received by the other set. The measurement end is where all the launch fibres and receiving fibres are combined, in a randomly distributed manner, across the polished end of the bundle. Such simple intensity-based sensors can monitor surfaces to resolutions of a micron or less, depending on the core sizes of the fibres, the nature and consistency of the reflecting surface they are monitoring, and the signal/noise ratio (SNR) of the optoelectronic detection system. This latter aspect of SNR is very important for many intensity-based sensors, and will now be quantified for simple systems.

Calculation of SNR of intensity-based sensors

It is useful to be able to calculate the SNR expected for intensity-based sensors. Most low frequency optical receivers used in sensors are based on PIN silicon diodes, followed by an FET transimpedance amplifier to convert the photocurrent to an output voltage. The noise in such receivers is usually limited by the Johnson noise in the feedback resistor, up to a point where the output voltage reaches about 50 mV, above which the receiver becomes shot-noise limited. Most practical intensity sensors operate in this shot-noise-limited regime, so we shall assume that this is the performance limitation here.

If we assume the mean optical signal power received at the detector is P_{det} and the fractional change in signal due to the effects of the measured parameter is $\mathit{\Delta}$ ($\mathit{\Delta}$ is an intensity modulation index):

> Then the desired modulation component in the optical signal due to these is $\Delta \cdot P_{\mathrm{det}}$ and the resulting modulation component in the detected photocurrent is $\mathit{\Delta}.R_{\mathrm{s}} \cdot P_{\mathrm{det}}$, where R_{s} is the responsivity of the detector (in $\mathrm{A.W}^{-1}$).
>
> The mean level of the detected photocurrent, I_{mean}, is given by $R_{\mathrm{s}} \cdot P_{\mathrm{det}}$.
>
> The shot noise in the photocurrent is given by the standard formula: $I_{\mathrm{n}} = (2 \cdot q.I_{\mathrm{mean}} \cdot B)^{-0.5}$, where q is the electronic charge and B is the post-detector-filter bandwidth of the signal.
>
> The SNR is therefore given by: $\mathit{\Delta}.R_{\mathrm{s}} \cdot P_{\mathrm{det}} \cdot (2 \cdot q.I_{\mathrm{mean}} \cdot B)^{-0.5}$.
>
> Manipulating, the $\mathrm{SNR} = \mathit{\Delta}.R_{\mathrm{s}} \cdot P_{\mathrm{det}} \cdot (2 \cdot q.R_{\mathrm{s}} \cdot P_{\mathrm{det}} \cdot B)^{-0.5}$.
>
> Hence: $\mathrm{SNR} = \mathit{\Delta} \cdot (R_{\mathrm{s}} \cdot P_{\mathrm{det}}/2 \cdot q.B)^{0.5}$.

The SNR therefore increases in proportion to the modulation index, $\mathit{\Delta}$, induced by the measurand, and also in proportion to the square root of detected signal power. Reducing the detection bandwidth, B, improves the SNR in proportion to the square root of the factor by which the bandwidth is reduced.

Problems with simple intensity-based sensors and the need for sensor referencing

The use of intensity-based sensors is very attractive, because of the ease of measuring the intensity (e.g. from the detected photocurrent in a silicon photodiode). However, simply measuring the output intensity from a single fibre containing a sensor can lead to a number of sources of error/uncertainty in the state of the sensor. The following factors can affect the output signal and hence cause errors in the observed sensor reading:

- light source variations;
- fibre lead and connector variations;
- variations in detector response;
- detector preamplifier noise.

In practice, the noise and variations in response of the optical receiver are not usually a major problem when using silicon detectors at low frequency. However, typical changes in light source intensity can be significant and changes in the transmission of optical fibres and connectors can be a real problem, particularly where there is mechanical deformation. In order to resolve the above problems, it is usually necessary, when making a quantitative measurement, to reference the sensor. This involves providing a reference optical signal, which, ideally, has not been changed by the measurand in the same manner as the sensing signal.

The simplest method of referencing is based on the provision of more than one fibre path from the light source, leading to two separate detectors, with the sensing section included in only one of the paths. The sensing signal is then normalized to the reference signal by division. More complex systems are possible, with routing of signals via separate paths using couplers [20] or fibre switches. Alternatively, fibre paths of different length can be used to separate signals from a pulsed source. Here the separation occurs in the time domain, according to the propagation delay each signal has experienced in its separate optical path.

In some cases, the need to measure the actual intensity can be avoided if the light is intensity modulated in a periodic way by the sensor head, and, rather than measuring the intensity, the quantity monitored is the frequency (rate) or time of modulation. Measurement of modulation frequency is not only easy, but it also avoids the need for sensor referencing, and the accuracy of measurement of cycles in a given period increases linearly with the measurement period, provided no modulation-cycle counts are missed, so very high accuracy is possible over extended periods. Very simple optical tachometers (figure C3.1.10) can be designed, using mechanical parts that give a periodic amplitude variation of light by virtue of variations in their light-blocking (e.g. a light-chopper wheel with radial slots) or light-reflective (e.g. a cylindrical metal shaft, with flats polished on it) properties as they rotate.

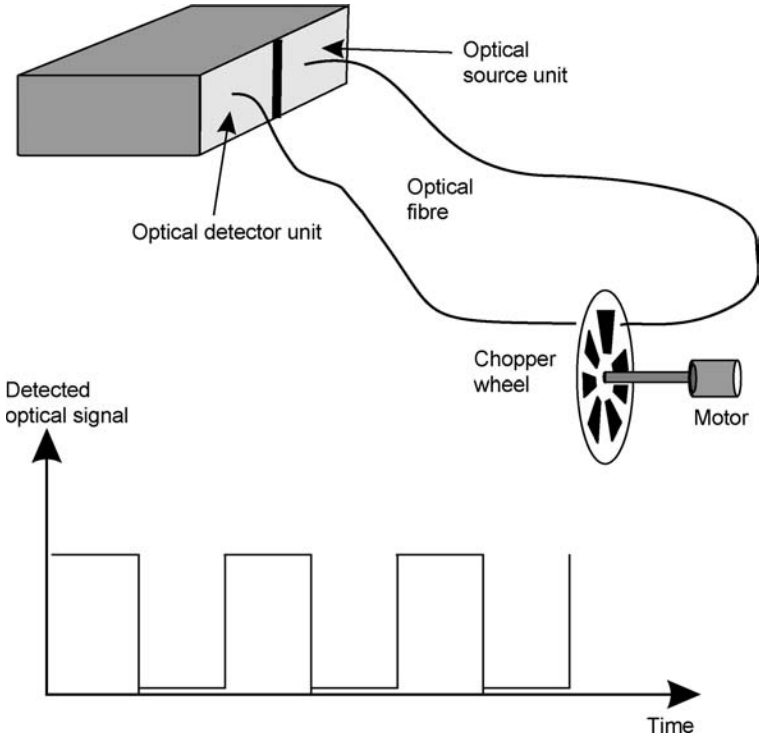


Figure C3.1.10. The optical fibre tachometer devised by M Johnson. The rate of beam interruption allows detection of shaft rotation speed.

Another attractive method uses silicon resonator technology, of the type now used in many modern electronic transducers. Here the modulator element is a tiny micro-machined silicon structure, such as a flexible cantilever or bridge that is driven into mechanical oscillation (in conventional sensors, by electrostatic or piezoelectric actuation) and its period is measured. Many physical influences (e.g. temperature, mechanical tension, pressure) can be arranged to interact and change the resonant frequency of such a structure, e.g. by changing the tension in the bridge element. The optical equivalent [21, 22] is essentially the same concept, but the element is driven by light exiting from an optical fibre, which is modulated to cause periodic heating of the microstructure (figure C3.1.11).

This forms a primitive heat engine, which gives rise to periodic thermal expansion and contraction of one side of the tiny silicon bridge (although photonic interactions have other means of causing mechanical changes in semiconductors), which drives this tiny element into resonant oscillation. The oscillation is of very low amplitude, unless the light modulation rate matches its resonant frequency.

The small motions of the oscillating element are also monitored optically, by changes in reflection amplitude (possibly involving simple optical interferometry to enhance the effects of small displacements) and the frequency of modulation is swept so the resonance can be detected, and even electronically locked onto. Such sensors appear at first sight to be a dream come true, seemingly solving the problems of measuring optical amplitude, and providing a 'frequency-out' measurement that has very high precision due to the high resonant frequency (typically ~100 kHz) of such microstructures. Unfortunately, there is a subtle problem, because the incident light heats the oscillating microstructure, and so changes its resonant frequency in a manner depending on the incident optical power. This gives the sensor an unfortunate degree of undesirable dependence on incident light intensity that still needs to be taken care of in design or operation of the sensor.

Spectral filtering sensors

A particularly attractive method of sensor referencing against intensity changes is to use spectral encoding in the sensor, where the sensor head provides an optical filtering function. Then, at the receiver, the relative signal strength at two or more optical wavelengths is monitored. If the sensor can filter out just one narrow spectral band, having a central wavelength dependent on the measurand, then this is

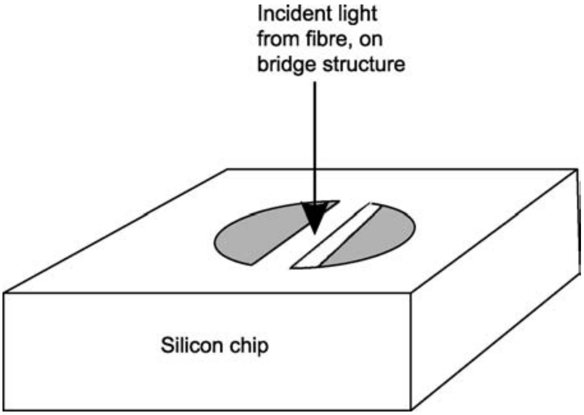


Figure C3.1.11. The silicon micro-resonator sensor, which is an optical equivalent of a sensor concept now common using conventional semiconductor technology. In the case of the optical sensor, excitation is by intensity-modulated optical radiation from the incident fibre. Detection of the resulting oscillations in the tiny silicon bridge can be performed by monitoring reflected intensity.

even more attractive, as this wavelength can then be monitored almost independently of the optical power level or intensity.

Intrinsic spectral filtering sensors using doped fibres

Rare-earth-doped fibres can change their spectral attenuation significantly with temperature, as the occupancy of internal electronic energy levels changes. These fibres can be used in short lengths or coils, as point temperature sensors, or simply as sensors designed to warn of fire or overheating affecting a section of what may be a much longer fibre.

Fibre grating sensors

Fibre grating sensors [23–25] use optically written structures in the light-guiding core of the optical fibre, which act as very narrow band optical filters (figure C3.1.12). These have a very narrow band reflection spectrum (typically between 0.1 and 1 nm bandwidth) and usually have a peak back-reflection of between 5 and 100%. In transmission, they have a similar bandwidth, but act as blocking filters over a similar narrow band.

The grating consists of a short region of fibre (typically 1–20 mm length, L), where the refractive index of the fibre core has been modified to cause it to exhibit a periodic variation with length in the axial direction. This variation is achieved by lateral illumination of a fibre with two converging beams of ultraviolet light (figure C3.1.12(*a*)), either created with beam-splitting optics or created by diffraction of a single beam in a phase mask. The beams interfere to give fringes, with a pre-designed spatial variation of optical intensity, depending on the wavelength and angle of convergence. This bright and dark fringe pattern gives rise to a corresponding refractive index variation in the photosensitive fibre core as a result of a *photo-refractive* effect. This effect is small, yet significant in germania-doped silica fibres. Even a very small periodic refractive index variation can build up coherently to cause a very significant (even close to 100%) reflection of light at the Bragg wavelength, λ_{Bragg}. This is the wavelength at which each low-intensity wavelet, reflected from each minor undulation in refractive index, can add coherently with all the others from other parts of the grating, and is given by:

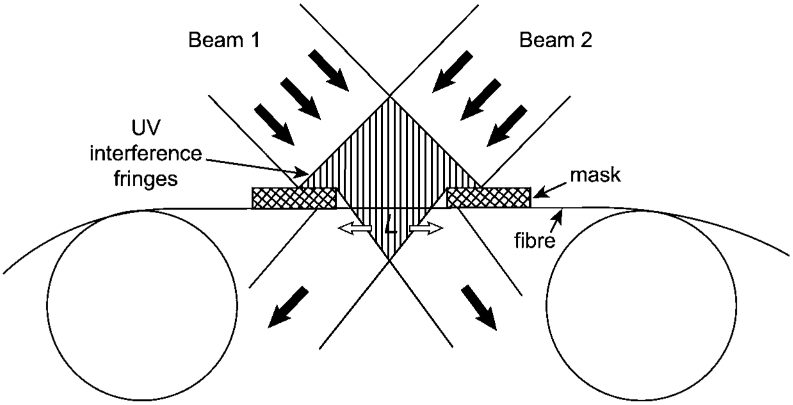
$$\lambda_{\mathrm{Bragg}} = 2{\cdot}n_{\mathrm{eff}}{\cdot}\Lambda$$

where n_{eff} is the effective refractive index of the fibre core and Λ is the grating period.

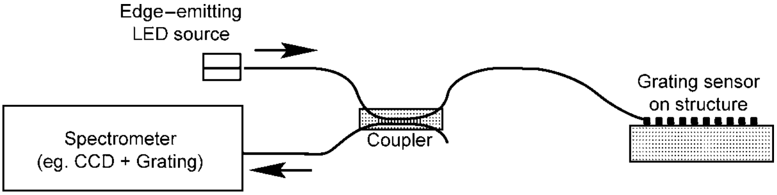
It is this need for this phase-matching, or constructive addition of back-reflected wavelets, in order to give significant reflection, that results in the desired narrowband filtering effect; the longer the grating, the narrower the filter effect and the higher the grating reflection coefficient for a given index change.

The in-fibre grating is an excellent sensing element [26] as it is tiny, and even without external components, can be configured to sense mechanical strain influences. As with traditional electrical resistance strain gauges, the wavelength of the grating depends not only on strain, but also on temperature, so there is a need to measure or compensate for temperature. The review on grating sensors by Kersey [27], and the textbooks in the bibliography cover many ways in which this may be done, but the simplest method is to have another (unstrained) grating as a reference thermometer. If it is merely required to monitor bending of a thin plate, then it is possible to bond identical gratings on opposite sides of the surface and simply monitor the differences in wavelength shifts observed [28].

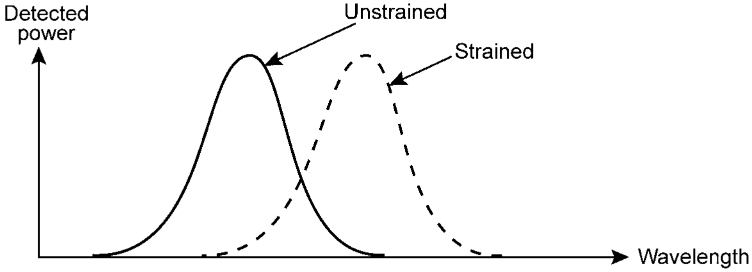
From [26] a typical thermal sensitivity of a grating with 1300 nm centre wavelength is of the order of $0.01\,\mathrm{nm}\,^{\circ}\mathrm{C}^{-1}$, with the fractional change per unit temperature being typically $\sim 8.2\times 10^{-6}\,^{\circ}\mathrm{C}^{-1}$ at all wavelengths. The fractional response of centre wavelength to strain is typically of the order of 75% of the strain.



(*a*) Inscription of grating in fibre, using converging uv light beams (grating is produced over exposed length *L*)



(*b*) System for the interrogation of grating wavelength, using spectrometer



(*c*) Response of system shown in above diagram, as grating wavelength changes

Figure C3.1.12. The in-fibre Bragg grating sensor. This shows (*a*) how a fibre grating can be written using interference between converging beams of light, (*b*) how a grating can be bonded onto the surface of a structure to be monitored (embedding into composite materials is also common) to form a strain sensor and (*c*) the response of the sensor as the structure is strained.

A major application area is in monitoring [29, 30] of composite materials (with the grating embedded in the composite) or of large mechanical or civil engineering [31] structures, with the gratings embedded in grooves or bonded to the surface (figure C3.1.12(*b*)) or attached in surface patches. With appropriate interrogation systems, and with mechanical amplifiers, grating sensors are even being considered for optical hydrophones to detect weak acoustic signals in the sea. Clearly, there are many physical parameters that can be arranged to give a strain in the fibre (e.g. hydrostatic pressure or depth, magnetic field). Although they are only optically written, the gratings have been shown to withstand elevated temperatures of several hundred degrees Celsius, but of course, the wavelength varies with temperature when heated or cooled.

In order to sense the grating wavelength, many interrogation methods [30, 32–36] have been devised. The most common method is to launch light from a broadband source (LED or fibre superluminescent source) into the fibre lead to the sensor and then detect the peak reflective wavelength of light reflected from the in-fibre grating, using some form of spectrometer (figure C3.1.12(*a*) and (*b*)). The most common types used are fixed diffraction-grating/CCD spectrometers, which usually require some form of fibre depolarizer to avoid errors [36] due to polarization effects, or scanned narrowband filters (e.g. using Fabry–Perot, as in [27] or acousto-optic tuneable filters, as in [35]. Systems where another grating is simply stretched to form a filter, that can be tuned to track the sensor grating, have been reported by Jackson *et al* [32]. There are also simpler passive methods, where the wavelength change is converted to an intensity change, using some form of optical filter having a high slope of transmission versus wavelength. Examples include interference filters and wavelength-selective fused fibre couplers [30, 150], of a type often used for separation of two closely spaced wavelengths. Ideally here, the light is split into two channels, one increasing in intensity, the other decreasing, as wavelength is increased, so the two complementary outputs can be ratioed to remove the common mode intensity changes that could occur due to variations in the light source, fibre leads or connectors.

Apart from the methods where the grating merely acts as a filter, it can also be used as sensor, in a configuration where it forms part of an active fibre laser [34]. There are two basic types here, the first being where a conventional in-fibre grating (the sensor grating) acts as one mirror of an optically pumped rare-earth-doped fibre laser and the other mirror is arranged to be broadband. As the grating is stretched (or heated), changes in its Bragg wavelength cause the laser output wavelength to change. The advantages are that the linewidth is much smaller, giving improved spectral and strain resolution, and the optical output power is much larger than for the alternative simpler configuration with filtering of a broadband source. A second, more compact, form of the same idea is achieved using in-fibre distributed feedback Bragg lasers (DFB fibre lasers) as the sensing element. Here the active rare-earth-doped lasing region forms part of the actual Bragg grating sensing element. The resolution of such sensors is nothing short of dramatic, as they oscillate at a frequency around 1014 Hz, yet they can be arranged to be monitored to 1 Hz resolution, using ‘mixing-down’ by beating or heterodyning with a reference laser, giving a potential resolution of 1 part in 1014. More usually, the beat frequency between two orthogonally polarized lasing modes has been monitored, as this is more easily used, being typically in the range of 0.5–2 GHz, where the beats can be monitored directly using a fast optical detector [37, 38]. These publications also discuss how thermal compensation is possible using monitoring of both the actual laser frequencies and the beat frequency between two mixed lasing modes.

Extrinsic spectral filter sensors

Many types of *extrinsic spectral filtering sensor* have been constructed [152]. One of the first methods [39] was to use a material which changes its spectral transmission characteristics with temperature (figure C3.1.13).

There are many examples of such materials (e.g. crystal, glass or polymer). As one attractive group of materials, most semiconductors have steep transmission band edges, beyond which they suddenly become reasonably transparent. At the *band edge* of all common semiconductors (e.g. Si, Ge, GaAs), both the transmission versus wavelength slope and the wavelength of 50% transmission vary significantly with temperature. There is also a large family of commercially available long-pass optical filter glasses (e.g. Schott and Corning glass companies) with semiconductor-like optical behaviour. Clearly, all of these can be used to make practical and stable sensors, and many low-cost spectrometers are now available to interrogate them. They are attractive for monitoring in remote or inaccessible locations, or in areas of high E-M field.

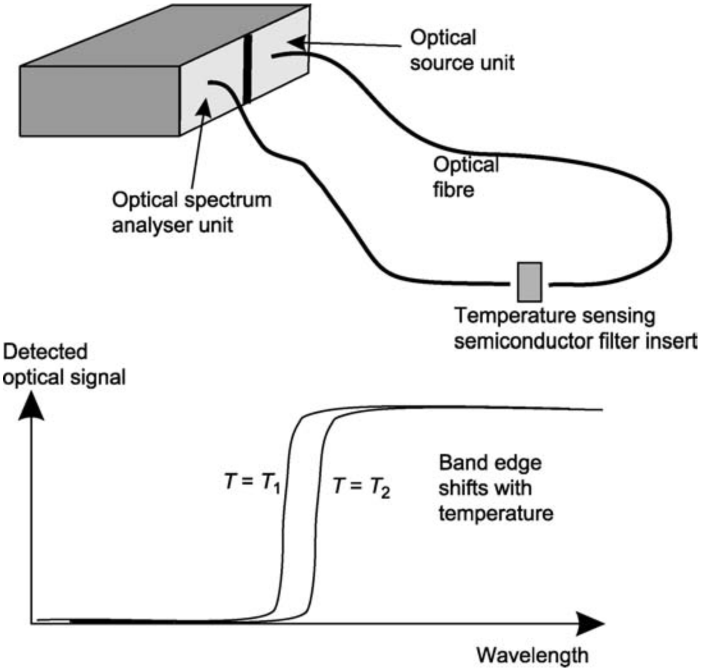

Optical source unit
Optical spectrum analyser unit
Optical fibre
Temperature sensing semiconductor filter insert
Detected optical signal
$T = T_1$
$T = T_2$
Band edge shifts with temperature
Wavelength

Figure C3.1.13. An optical fibre thermometer, where the thermal dependence of the band-edge position of a semiconductor (or special filter glass with semi-conductor-like properties) is monitored with a simple spectrometer.

A class of spectral-filtering sensors is based on fibre-coupled versions of common bulk optical spectrometer components, such as Fabry–Perot filters (closely spaced mirrors), diffraction grating, zone-plate, or prism spectrometers or monochromators. The Fabry–Perot filter is particularly useful, in view of its small size and narrow linewidth, and the ability to coat fibre end surfaces to make a tiny version from a short fibre section with reflective end coatings [40].

Extrinsic sensors using monitoring of fluorescence spectrum

Apart from using simple transmission measurement, the fluorescence spectrum of some materials can be monitored, a method again well suited to the use of direct-bandgap semiconductor crystals [41] in the sensor probe. However, in this mode, there is now the possibility of using translucent materials, such as the types of phosphors used in fluorescent lights and in TV display tubes, which, with the aid of a directional coupler, can be conveniently monitored in back-'scatter' mode via a single fibre lead. This fibre-coupled configuration (see later figure C3.1.15(*b*)) will be described again, when we discuss chemical sensors.

Decay lifetime sensors can sense various parameters by monitoring the time delay after excitation of the sensor probe material. Excitation is either by a short optical pulse or by a repetitive pulse train, or using modulated light having some other (e.g. sinusoidal) periodic intensity variation. The most common mechanism is to look, directly or indirectly, at the time decay of fluorescence in the probe (figure C3.1.14).

The fluorescence decay curve, which is usually of an exponential shape, can be monitored, to determine the lifetime. Alternatively, the phase delay in the detected fluorescence intensity signal, relative to the initial intensity modulation waveform (usually sinusoidal or square-wave) of the incident light, can be monitored.

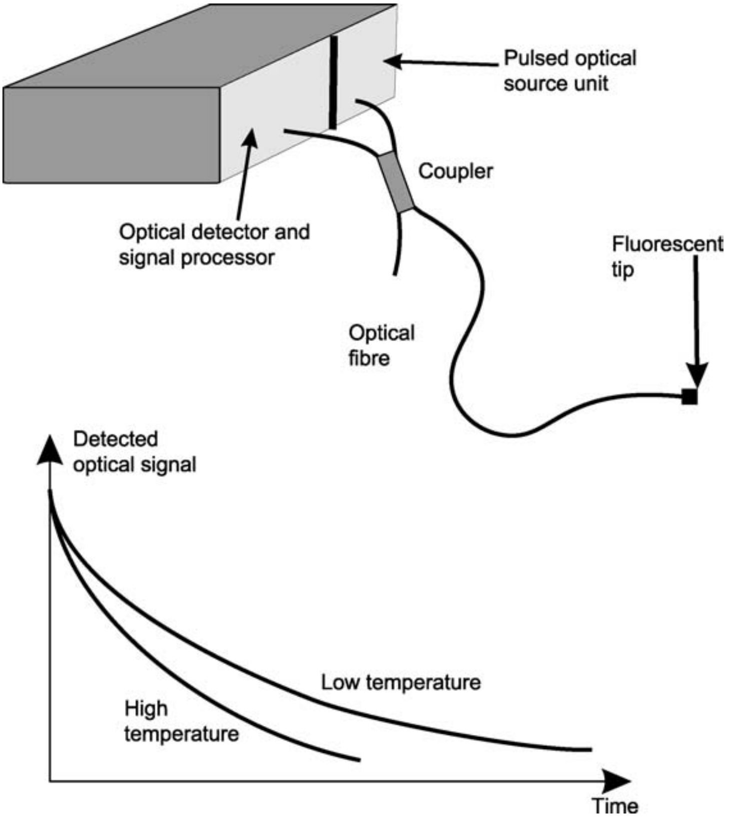
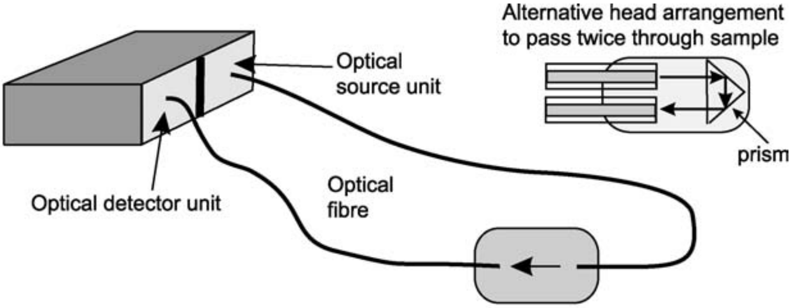


Figure C3.1.14. Sensor to measure temperature, where the decay of fluorescence of a phosphor (or semiconductor chip), attached on the end of an optical fibre, is monitored. The lower curves show typical responses, following pulsed excitation from the light source.

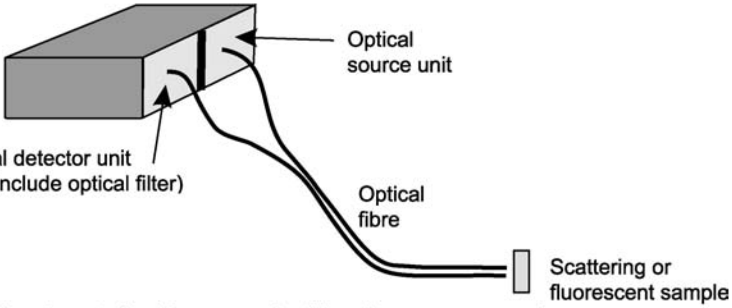
Several practical physical and chemical sensors use this mechanism, and it has proved particularly useful for temperature monitoring. Thermal changes affect the fluorescent lifetime of many substances (e.g. phosphors, semiconductors, laser crystals or glasses), but phosphors [42] and ruby crystals [43] have been used most for thermal sensing (see also review by Gratton and Palmer [44]). Many of these materials have been used extensively for long periods under conditions of energetic electron bombardment or powerful optical illumination, and have excellent long-term stability. Laser crystals, in particular, are not only optically and thermally stable, but also often (e.g. ruby) have excellent mechanical hardness and strength.

C3.1.2.2 Intensity-based chemical sensors

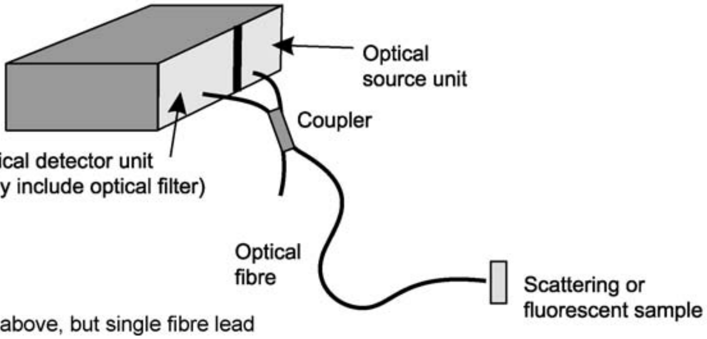
There are two basic types of optical chemical sensor, one using *direct optical spectroscopy* of materials to be detected, the other making use of a *chemical indicator*, i.e. a compound that acts as an intermediary, with a strong, hopefully chemical-species-specific, change in its optical properties when exposed to a target chemical or group of chemicals. Below, we shall cover these two types, but, in this short outline, it will not be possible to cover all the concepts concerned, so the interested reader is referred to the textbook in bibliography 1 (particularly vol 4, chapters 7 and 8) for further reading. For reasons of lack of space, the discussion of chemical sensors here will be shorter than that for physical sensors.


Alternative head arrangement
to pass twice through sample
Optical
source unit
prism
Optical detector unit
Optical
fibre

(*a*) Transmission spectroscopy
(interrogator can be conventional spectrophotometer)


Optical
source unit
Optical detector unit
(may include optical filter)
Optical
fibre
Scattering or
fluorescent sample

(*b*) Sensing of scattering (Raman or elastic) or fluorescence sample


Optical
source unit
Coupler
Optical detector unit
(may include optical filter)
Optical
fibre
Scattering or
fluorescent sample

(*c*) As above, but single fibre lead

Figure C3.1.15. Optical fibre sensors for remote spectroscopy. This shows (*a*) transmission spectroscopy, where, both the source and interrogator unit can be an integral part of a commercial spectrophotometer, modified to allow fibre extension leads; (*b*) sensing of backscattered signals, where, a filter or spectrometer can be used to select out Raman or fluorescent signals from the incident light if desired; and (*c*) a modification to allow a single fibre lead to the measurement head.

Chemical sensors using direct optical spectroscopy

Because materials can absorb or emit light only at wavelengths corresponding to allowed internal energy level changes, characteristic of particular elements or functional groups, spectroscopy has become one of the most valuable tools of the analytical chemist. Most of the standard spectroscopic techniques used by analytical instrument designers can also be implemented in fibre optic form (see examples in figure C3.1.15).

Possible methods include measurement of transmission and turbidity, attenuated total reflection, fluorescence, Raman scattering (including surface-enhanced Raman scattering, or SERS), to give just a few examples. The great advantage of fibres is the real-time, online measurement capability, allowing the instrumentation to be kept in a benign environment, away from the probe end, that may be remote or inaccessible, and could have any variety of dangerous, corrosive, toxic or flammable materials present.

Transmission (absorption) and turbidity measurements

Transmission defines the fraction of light passing through a component or medium. Absorption is the total loss of light arising from conversion of light to heat, whereas turbidity is where light is merely lost from a collimated path by scattering, such that it is no longer collected.

The power, $P(\lambda)$, transmitted through a sample in a small wavelength interval at a centre wavelength λ, is given by Lambert's law:

$$P(\lambda) = P_0(\lambda) \cdot \exp[-\alpha(\lambda) \cdot \ell]$$

where $P_0(\lambda)$ is the power entering the sample in this wavelength interval, $\alpha(\lambda)$ is the attenuation coefficient of the material at wavelength λ and ℓ is the optical path-length through the sample to the point at which $P(\lambda)$ is measured. Lambert's law does not apply if scattering is high, such that multiple scattering occurs.

The sample can be said to have a transmission $T(\lambda)$, at the wavelength λ, where:

$$T(\lambda) = \exp[-\alpha(\lambda) \cdot \ell].$$

Alternatively, the sample can be said to have an absorbance $A(\lambda)$, where:

$$A(\lambda) = \log_{10}[1/T(\lambda)] = \log_{10}[P_0(\lambda)/P(\lambda)] = 0.43\alpha(\lambda) \cdot \ell.$$

The factor 0.43, or log(e), has to be included to account for the use of $\log_{10}$ for absorbance calculations, whereas natural exponents are used for attenuation coefficients.

Transmission, absorption or turbidity measurements can be achieved most easily by using a commercial spectrophotometer with extension leads. These have a unit that fits into the cell compartment of a standard instrument, with a first lens that takes the collimated light that would normally pass through the sample chamber, focuses it instead into a large core diameter (usually $>200\ \mu$m) optical fibre down-lead and with a second lens that re-collimates light coming back from the return fibre lead into a low-divergence beam suitable for passage back into the instrument. There is also a remote measurement cell, connected to the remote end of both these fibre leads, where a first lens collimates light coming from the down-lead into an interrogation beam, which passes through the remote measurement cell, after which a second lens collects the light and refocuses it into the fibre return lead going to the spectrometer instrument. Such optical transformations lead to inevitable losses of optical power, of typically 10–20 dB (equivalent to losing 1–2 units of absorbance), but as most modern instruments have a typical dynamic range of >50 dB, this is a price that many users are prepared to pay for a useful remote measurement capability.

It should be noted that the optical power losses occur mainly due to the imperfections of the focusing and re-collimation optics and due to reflection losses at interfaces, rather than to fibre transmission losses. If suitably collimated beams were to be available in the instrument, if large core diameter fibres could be used to connect to and from the probe, and if all optics, including fibre ends, could be anti-reflection coated, there should really be very little loss penalty. Such losses therefore arise primarily because of the need for the fibre leads to be as flexible as possible (so hence choice of small diameter fibres) and the usual need to compromise design on grounds of cost.

There are many other probe head designs that are possible. The simplest design, for use with measurement samples showing very strong absorption, is simply to have a probe which holds the ends of the down-lead and return fibre in axial alignment, facing each other across a small measurement gap, where the sample is then allowed to enter. Losses are low for fibre end spacing of the same order as the fibre diameter or less, but rapidly increase with larger gaps. The probe is far easier to miniaturize and handle if the fibre down-lead and return lead are parallel in one cable. This can be achieved using a right-angled prism or other retro-reflecting device to deflect the beam in the probe tip through the desired 180° that allows it to first leave the outgoing fibre, pass through a sample and then enter the return fibre. Use of a directional fibre coupler at the instrument end allows use of a single fibre, but then any residual retro-reflection from the fibre end will be present as a crosstalk signal, adding light signal components that have not passed through the medium. Clearly, there are many variants of such optical probes, some involving more complex optics (e.g. multi-pass probes), some constructed from more exotic materials to withstand corrosive chemicals.

A very simple option that has often been used with such single fibre probes, for monitoring the transmission of chemical indicators, is to dissolve the indicator in a polymer that is permeable to the chemical to be detected, and also incorporate strongly scattering particles in the polymer. When a small piece of such a polymer is formed on the fibre end, the particles give rise to strong backscattered light, and the return fibre guides a portion of this to the detection system. This backscattered light had of course to pass through the indicator polymer in its path to and from each scattering particle, so the returning light is subject to spectral filtering by the indicator. Although this is a very lossy arrangement, it is extremely cheap and simple, and has formed the basis of many chemical sensors, for example ones using pH indicators [45].

There are now many commercial types of miniature CCD spectrometers that have been specially designed to analyse the spectrum exiting from an optical fibre. These generally use the light-guiding core of the return fibre as the input ‘slit’ of a diffraction grating monochromator, using a sensitive CCD detector array to provide a set of parallel output signals, one from each narrow-band spectral component of the received light. These have high optical efficiency, as nearly all the energy is incident on the detector. For ultimate resolution, however, the well-known Fourier transform spectrometer principle can be used, where a scanned optical interferometer is used to analyse the spectrum of the received light. Each optical spectral component gives its own electrical frequency component in the detected photocurrent signal, observed as the interferometer is scanned (a narrowband laser signal input will give a pure sinusoidal output from the scanned interferometer). The entire spectrum can therefore be obtained, by Fourier analysis of the temporal variations of the detected signal, to extract the relative magnitude of each frequency component in this detected signal.

Chemical sensing by detection of fluorescence and Raman scattering

It was stated above that, for transmission and turbidity measurements, there should really be only a small additional loss penalty due to using optical fibres, provided full use is made of expensive precautions to avoid loss of light. This is unfortunately not true when fluorescence or Raman measurements are required, as light is scattered over all angles, and the return fibre can only ever collect a relatively small portion of this light. In addition, the processes where this *inelastically scattered* (i.e. different wavelength to that incident) light is generated, often have a poor *quantum efficiency* (i.e. have a low ratio of total re-emitted to total incident photon flux).

An arrangement similar to that of figure C3.1.15(*b*) has been used for oxygen sensing, using Ruthenium dye complexes [46], where the fluorescence decay is quenched by the oxygen gas.

There is no space here for a full discussion of these processes, but their potential value, particularly that of Raman scattering, for chemical sensing, has meant workers will continue to persevere to get useful performance, despite the low return light levels encountered with fibre-coupled systems.

Both these mechanisms involve excitation of a sample with light, usually at a wavelength shorter[†] than the scattered light to be observed, and then the re-emitted light is collected and narrow-band filtered. This filtering is firstly to separate it from the incident light, but also, in the case of Raman, to examine it for the spectral features characteristic of a target compound to be examined.

It is useful to briefly estimate the approximate magnitude of additional losses when using fibres with these processes. The loss due to launching of excitation light into a fibre is usually negligible with Raman, as powerful narrow-line laser sources are used, but ultimately the limit may be set by nonlinear processes or, in the case of large-core multimode fibres, by optical damage thresholds. Similar excitation can be used for low-level fluorescence monitoring, provided no photo-bleaching or other photo-degradation of the monitored substance can occur at high illumination intensity. The main potential loss is therefore that of light collection. If we assume the medium is excited close to the end of a fibre, only a region of dimensions of the same order as the fibre core diameter will be intensely excited. Thus, if a very large 200 μm core diameter fibre is used, a region of approximately 200 to >500 μm in length will be excited effectively. As a re-emitting point in the excited medium gets further from the fibre tip, the brightness of its illumination reduces and the effective angle subtended by the fibre core collecting the light (which can be approximately considered to be re-emitted in all directions) gets smaller. The collection can therefore be approximated as that received by the fibre numerical aperture (say NA of 0.3) from a region around 200 μm long. In a normal Raman spectrometer, the sample can be excited by a narrow focused laser beam. There can therefore be a useful and collectable Raman emission from a thin sample region, of length equal to that of the spectrometer input slit on which it must be focussed. The length of the slit may be perhaps 10 mm long, and light entering this may perhaps be collected with a wide acceptance angle monochromator, perhaps having an NA as high as 0.4. Thus, the fibre-based system may perhaps have a light collection reduced by a factor of $50 \times 16/9$ (product of the useful excited length ratio and the square of the acceptance NA ratio), or approximately 90 times, when compared to a bulk optical system. A similar factor also applies to fluorescence detection.

Apart from these photometric limitations, Raman scattering has a particularly poor quantum efficiency, and the already weak scattered signals will be typically two orders of magnitude weaker when coupled into return fibres, with their poor collection efficiency for divergent light (see bibliography 1, chapters 7 and 13). Despite this, however, several commercial fibre-coupled Raman systems are available. These combine the great advantage of Raman scattering (which, merely by careful spectral filtering, allows rejection of elastically scattered light from turbid samples) with the ability of fibres to probe into inaccessible, remote or hazardous environments.

A few practical examples of chemical sensing using direct spectroscopy

There are many more means of performing direct spectroscopy with optical fibres, as the above has only presented a few. Other methods will be discussed in the later section on distributed and multiplexed sensing. As space is limited here, it is instructive to summarize, in the form of a brief list of examples, where systems have been built for practical applications. In many cases, active research is still being pursued in many of these areas:

- Sensors examining transmission of groundwater to track pollution.
- Fibre-probe sensors to examine the transmission or reflection spectrum of blood, to determine oxygenation state.

[†] Anti-Stokes Raman light, at wavelengths shorter than the incident light, although normally very weak, can still have a useful intensity if it is measured at a wavelength very close to the incident wavelength. It then often has less interference from fluorescence light.

- Fibre-probe sensors to examine bilirubin in the digestive system of the body.
- Gas sensors, based on remote absorption measurements, using fibre probes.
- Sensors examining the transmission of petrochemicals to determine octane rating. These can be extrinsic sensors as in figure C3.1.15(*a*) or can use evanescent field monitoring.
- Fluorescence sensors to determine oil or other aromatic hydrocarbons in water.
- Technologies using arrays of optically addressed 'micro-dots', each dot in the array having a different composition. For example, arrays of fluorescent indicator dots, each dot having different optical properties, can detect several different substances using one fibre (or can give multiple signals to cancel crosstalk from other substances, using mathematical regression models).
- Methods using evanescent wave coupling to the measured substance, for example, with a bare glass or silica fibre. Alternatively, a conventional monomode fibre with part of the cladding polished away, or a 'D-type' fibre that has a very thin cladding on one side, can be used. In such sensors, the field of the guided light extends beyond the vitreous waveguide into the chemical to be measured, allowing absorption of the latter to be observed as fibre attenuation.
- Technologies as above, but using *surface plasmon resonance* to enhance the coupling to the evanescent field.
- Refractive index sensors, where a fibre with Bragg gratings is side-polished, removing cladding material to allow the evanescent field to extend outside the fibre. Then, the peak reflected wavelength of the grating is affected by the effective propagation constant of the fibre, which is now also a function of the refractive index of this surrounding medium. Again, octane rating of fuels is of interest here.

The textbooks in bibliography 2 give a very comprehensive overview of many of these types of optical chemical sensors.

C3.1.3 Interferometric sensors

We shall now discuss interferometric sensors. Here, the sensing action involves the interaction of fibre-guided light beams, where there is coherent addition of the electric field components of their electromagnetic waves. This leads to a mixing condition that can vary from constructive to destructive interference, depending on the relative phase of the combining light beams. In order to obtain high visibility interference of interfering free-space beams, it is necessary to match their intensity profiles, their wavefront shapes (i.e. beam direction and divergence) and their polarization states over the full transverse width of the interfering beams. In order to observe or detect the effects of interference, however, the light beams must also eventually fall onto a 'square law' optical detector. All standard optical detectors, including the human eye, monitor optical power or intensity, which is proportional to the square of the electrical field component, hence the term 'square law'.

C3.1.3.1 Interferometers using fibre optic technology

A single mode optical fibre guides not only one propagation mode, but also strictly two, if the two possible orthogonally polarized modes are considered. The fundamental guided mode is usually called the HE_{11} mode. This fibre mode has quite a simple transverse power distribution, very closely matching the well-known Gaussian TE_{00} intensity profile, typical of that of the beam from a single transverse mode gas laser, with a central (axial) peak in the intensity. The big advantage of monomode fibres is that (apart from the possibility of the two different principal polarization modes) only a single spatial mode is allowed, the fibre acting as a perfect mode filter to ensure that the two overlapping fields have identical spatial characteristics. The slight curvature of E-M field lines in the fibre, varying a little with transverse position, does not significantly reduce fringe visibility, as the other interfering signal (or signals) guided in the same fibre have a matching curvature.

Because of this behaviour, we can fabricate interferometers using single mode fibre to define the optical paths, and using compact fibre coupler components as beam splitters or combiners [47]. In figure C3.1.16(*a*) and (*b*), we show a Michelson and a Mach–Zehnder type interferometer, respectively, both implemented with single mode fibres.

Unfortunately, there is a big disadvantage when light is guided by an ordinary single-mode telecommunications-type fibre, as, unlike the situation for free-space, or in-air, beams, the polarization direction of the HE_{11} mode can easily be changed by environmental influences. For these to change the polarization of fibre-guided light, they need only cause a significant asymmetrical physical distortion of the fibre. Such influences can occur due to, for example, fibre bends, lateral mechanical stresses and transverse thermal gradients. In the extreme case of this so-called *polarization fading*, the visibility of the interference can fade completely, as orthogonally polarized guided beams cannot interfere. This is simply because electric fields at 90° can no longer cancel, the resultant intensity now being independent of optical phase of the combining beams.

To compensate for possible signal fading due to this polarization fluctuation, it is common to use polarization controllers (PCs) when using ordinary single mode fibres. However, in real-world sensors, it is not attractive having to have to continually adjust polarization in order to compensate for environmental effects, unless the PCs are themselves controlled by automated optoelectronic feedback systems. Although the latter is possible, it is still a rather complex and expensive solution. Fortunately, it is possible to greatly reduce such effects with passive solutions. One method is to use polarization-diversity

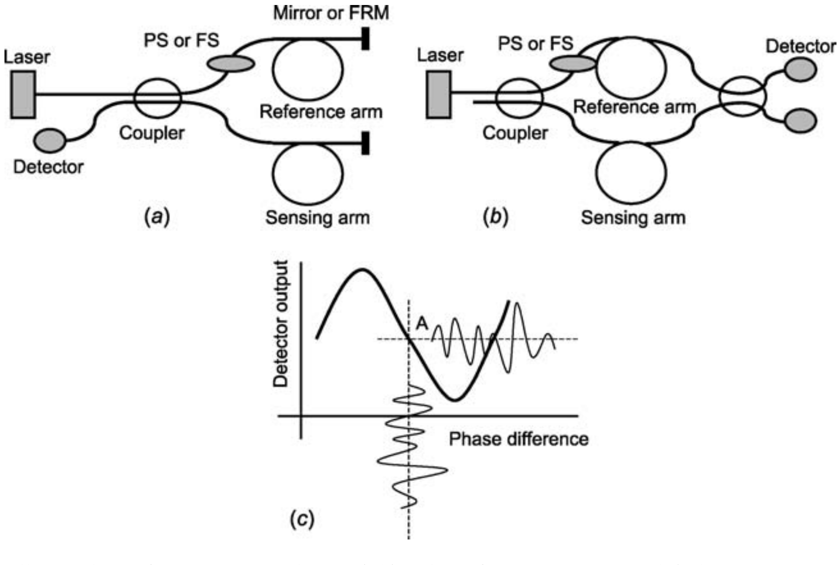


Figure C3.1.16. Illustration of two commonly used fibre interferometers: (*a*) Michelson arrangement, (*b*) Mach–Zehnder arrangement, and (*c*) the typical detector response to phase difference. Note that the zero position on the phase-difference scale, where intensity would be at a peak, is unknown in this figure.

optical receivers, where the combined light is first split into differently polarized components before detection on several separate detectors. A simpler and far more efficient passive solution, however, is to use polarization-maintaining fibre (P-M fibre). In this fibre, the direction of the linear polarization of the HE_{11} mode is constrained to lie along one of its principal polarization axes, as the lightwave propagates through it. Such a fibre is made with a deliberate transverse asymmetry, such that either of the two principal polarization modes, once launched, cannot readily couple to the other, unless there is a very severe localized deformation of the fibre. An even more effective method is to use single polarization fibres, which are polarization-maintaining types where one polarization mode is made to be highly lossy, such that only the other mode can effectively propagate without high attenuation. If any small degree of polarization mode conversion then takes place in this fibre, the light in the undesired mode is rapidly lost.

In the case of a dual-reflective path, optical fibre Michelson interferometer, we have another ingenious way to stabilize for any in-fibre polarization fluctuation. To achieve this, Faraday rotator mirrors (FRM) are placed at the end of the fibres. A Faraday mirror device has the property of ensuring the reflected lightwave has a returning state of polarization (SOP) rotated by 90° with respect to the incident wave. Consequently, after a two-way propagation along the fibre, the SOP of the lightwave always remains orthogonal to the incident one, even though the actual SOP of each beam along the fibre will always fluctuate. If a wave is split, by a coupler, into two arms, each having Faraday mirrors at their far (distal) ends, *both* returning beams or guided modes return with orthogonal polarization states to that of the launched light, so *both* have identical polarization states.

In a fibre interferometer, the detected output, I_d, is given simply by:

$$I_d = A + B \cos \Delta\theta$$

where A and B are constants and $\Delta\theta$ is the phase difference between the two lightwaves returning from the reference and the sensing arms. The output intensity changes, as a function of phase difference, $\Delta\theta$, are shown schematically in figure C3.1.16(*c*). In figure C3.1.16(*a*) or (*b*), a relatively large phase drift can always be caused by temperature or mechanical fluctuation, which results in the presence of low-frequency components, or drift, in $\Delta\theta$.

For many sensors, such as acoustic wave (e.g. fibre hydrophones) or vibration monitors, it is desired to monitor or measure very tiny phase changes, usually of a cyclic or transiently varying nature, caused by the external measurand, typically having a low- to mid-AC frequency content in the range 10 Hz to 10 KHz. In order to measure these small, relatively rapidly changing signals, other sources of slower drift, for example due to slow thermal changes or due to slowly varying mechanical strain, must be corrected for or stabilized, so as to keep the interferometer at its most sensitive operating point. This is the maximum slope of the sinusoidal response, termed the *quadrature* point.

To achieve the desired stabilization, an optical phase shifter (PS) can be placed in the reference arm, as shown in figure C3.1.16(*a*) and (*b*). In these stabilized interferometers, the detector output is fed-back to the PS, so that the output signal is held at *quadrature* point A, where the slope of the curve is steepest. In this way, tiny alternating optical phase signals, as small as 10^{-6} radians or less, can be detected. At an optical wavelength of 1 μm, this represents an optical path length change of only 10^{-12} m, corresponding to a change in the physical length of the sensing fibre by a tiny amount, corresponding to a small fraction of the diameter of a hydrogen atom! (Of course, individual atoms at the fibre ends have positional uncertainties greater than this, but the end position is the average of enormous numbers of such atoms.) This means that fibre interferometric sensors can realize extremely high sensitivity, and this sensitivity can be further increased by using a multi-turn fibre coil as a sensing element.

For many applications, it is convenient to place an optical frequency shifter (FS) in one arm, to deliberately induce a frequency difference between the two lightwaves. This is also shown in figure C3.1.16. Under these conditions, the detected output, I_d, can be expressed as:

$$I_d = A + B\cos(2\pi\Delta f t + \Delta\theta).$$

This new configuration is called a *heterodyne* interferometer, a term which implies that the interfering lightwaves now have different frequencies. The phase of the detected electrical signal can now be detected by conventional phase or frequency demodulation schemes, of a type commonly used in commercial or domestic radio receivers. One such electrical frequency demodulation method is to delay one signal compared to the other and use an electronic multiplier or mixer, followed by a low-pass filter, an arrangement that conveniently converts the electronic phase difference of two electronic signals to amplitude, with a linear-saw-tooth response characteristic. Another common method, again well known to radio engineers, is to use phase-lock loop demodulation technology. In order to distinguish the earlier, rather simpler, interferometer (in which the two lightwaves had the same frequency), from the *heterodyne*, or difference frequency one, we have just discussed, the earlier one has been termed a *homodyne* interferometer.

In the interferometer shown in figure C3.1.16(*a*) and (*b*), the optical angle-modulation elements, i.e. the PS or FS modulators, are placed in the reference arm, which is sometimes located adjacent to or near the sensing arm to assist with thermal compensation. If the electronically driven PS or FS devices are placed in, or near, the physical measuring environment, they might pick up electrical interference, which would reduce the advantage that an optical fibre sensor is normally not affected by, nor induces electromagnetic noise. This would remove the inherent *electromagnetic compatibility* of the sensor. To improve the electromagnetic behaviour, a configuration shown in figure C3.1.17 has been proposed, where, rather than using a separate modulator element, a frequency modulation is created in the laser light source, which then, due to different optical time delays, creates a phase difference $\Delta\theta_1$, between lightwaves that have travelled over different length optical paths.

This then gives a phase difference expressed by:

$$\Delta\theta_1 = \frac{2\pi}{c}\Delta L{\cdot}\Delta f_1$$

where ΔL is the optical path length difference between the two arms. The lasing frequency of a semiconductor diode laser can conveniently be changed merely by modulating the drive or injection

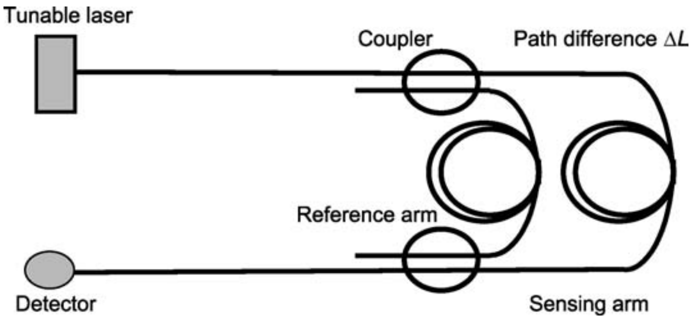


Figure C3.1.17. Configuration of a passive remote Mach–Zehnder interferometer. This allows remote interrogation by sweeping the laser source, avoiding the need for modulators with electrical drive signals to be included in the remote interferometer section.

current, providing a very simple current-controlled optical oscillator. By using this property, we can compensate for the effects of slow thermally or mechanically induced drift in the interferometer, without the need for any separate PS or FS. Using this configuration, there is then no need for any electronic or electrical components in the sensing environment, giving excellent freedom from electrical interference, etc. When the laser frequency is modulated by a sinusoidal wave of frequency ω_{p}, we induce a set of components at frequencies $m \times \omega_{\mathrm{p}}$ in the detector output, where m is an integer. Each of these components has a phase corresponding to the optical phase difference to be measured. This configuration shows, in many ways, a similar output to that of the heterodyne interferometer, but because no separate modulator is used, it is called the *phase-induced carrier method*.

Because of the over-riding need to reduce polarization fluctuation and phase drift in highly sensitive fibre optic interferometer sensors, many different methods have been proposed and developed by researchers. The textbooks in the bibliography present greater details of many of these schemes, but below we will present a case study of the optical fibre gyroscope to introduce some of the methods that can be employed. The gyroscope is a very special case, where it is essential to achieve a very high degree of steady-state stability. Fortunately, many other types of interferometric fibre sensor are designed to measure only dynamic changes, such as might arise from acoustic signals (e.g. the fibre hydrophone) or from other mechanical vibrations (e.g. the fibre vibrometer or accelerometer), so these do not suffer from quite so many undesired sensitivities to slow drift from environmental aspects as the fibre gyroscope can. The discussion below illustrates that, despite many fairly complex difficulties, a host of potential problems can be overcome to make a cost-effective practical sensor, although a great deal of background research was necessary to reach this point, and the gyro still needs careful design.

C3.1.3.2 High sensitivity sensing with fibre interferometers

The fibre optic gyroscope

The fibre optic gyroscope [48–50], which we shall call 'FOG' for short, was one of the earliest types of interferometric fibre sensor, and is the one which has perhaps received the most research funding and scientific attention.

A FOG detects rotation relative to an inertial frame. The basic operating principle of this sensor is based on a concept known as the 'Sagnac effect' [51], which originally used two optical beams, each directed in opposite directions around loops using mirrors, before being caused to interfere on a detector. The basic configuration of the all-fibre version is shown in figure C3.1.18.

It can be considered that two lightwaves, propagating in opposite directions in the same closed optical fibre coil, exhibit a travelling-time difference, which is proportional to the rotation rate of the optical path with respect to the inertial frame. This time difference results in a phase difference, θ, between the beams at the output of the loop, given by:

$$\theta = \frac{4\pi L a}{c\lambda}\Omega$$

where L, a, c, λ and Ω are, respectively, the length of the fibre coil, its radius, the speed of light in vacuum, the optical wavelength and the coil rotation rate. The magnitude of the Sagnac phase shift is generally extremely low for most typical rotation rates, particularly for those typical of vehicle navigation, where directional changes are usually quite slow. To overcome this limitation, a very long (and hence necessarily low loss) fibre coil needs to be used for the sensing coil. For aircraft navigation, a rotation rate resolution of only 0.01 $\mathrm{deg\,h^{-1}}$ is required. Even when using a sensing fibre as long as 1 km, this typically corresponds to an induced phase difference as small as 1 μrad. To measure such small and slow-changing phase changes, a successful FOG requires extremely careful control of the many subtle

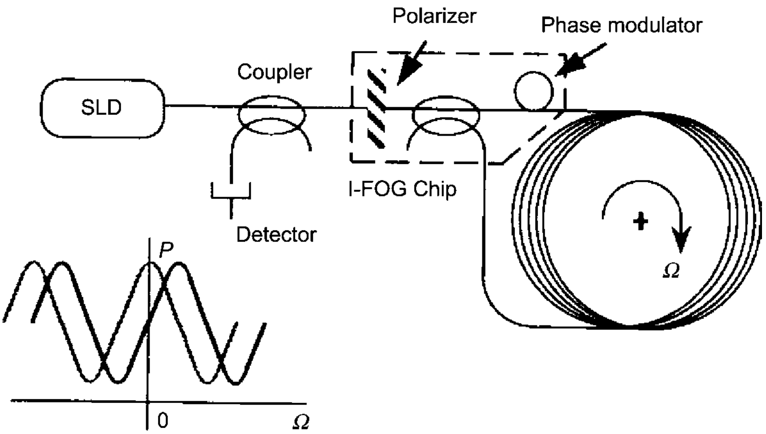


Figure C3.1.18. Basic configuration for a fibre optic gyroscope (FOG). This shows the phase modulation elements used to bias the device contained on an integrated optics chip (I-FOG chip) that also contains the polarizer and a beam-splitting element.

noise and drift factors that can otherwise occur. Fortunately, after extensive research over many years, these noise factors have been studied in detail, and many excellent countermeasures to reduce them have already been invented.

As a result of this success, FOGs are now feasible for many new applications of gyroscopes, such as car navigation, antenna/camera stabilizers, radio-controlled vehicles, unmanned track navigation and so on. These new applications could only be realized because of unique advantages of the FOGs compared to traditional mechanical gyros, such as short warm-up time (no need to build up mechanical speed, as in normal types), less maintenance and low cost. FOGs have also now been used in traditional application fields, such as airplane navigation, rocket launching and ship navigation. A part of the navigation system of the Boeing 777 now uses FOGs.

To measure tiny phase differences, of the order of micro-radians, between the two lightwaves travelling in the fibre coil of figure C3.1.18, a special processing scheme had to be established. When the system is in the rest state, the signals interfering at the detector in figure C3.1.18 would normally give rise to a maximum in the sinusoidal response (dashed curve). At this peak the gradient is of course zero, so there is no initial response to small phase changes, so hence zero sensitivity for a small rotation rate. In order to solve this problem, a phase biasing scheme is needed, in order to drive the state of the interferometer to a region of the phase response curve where the gradient is non-zero. To achieve this, a phase modulator, driven by a signal of sinusoidal or square waveform, is applied at, or near, the end of the sensing fibre coil. Due to the timing difference of the modulation between the clockwise (CW) and the counter-clockwise (CCW) waves, the two lightwaves now have a periodic phase difference when they impinge on the detector, and give a ‘mixed’ or detected output signal having the same frequency as the applied phase-modulation waveform.

By detecting the electronic output synchronously (i.e. by multiplying it with an electronic reference signal at the same frequency as the applied phase modulation, and then low-pass filtering), the signal shown by the solid curve can be obtained. Now it can be seen that the detected output changes, as desired, with input rotation. The polarizer shown in figure C3.1.18 is required to reduce drift effects due to the polarization fluctuation in the sensing fibre coil. Essentially, it ensures that light travels in each direction in the loop in different directions, but in the same polarization state, although, of course, each individual beam still exhibits polarization changes as it propagates round the loop. The system shown in figure C3.1.18 is called the 'minimum reciprocal configuration' for the FOG.

Even when the minimum configuration is used, many noise, drift and signal-fading factors can still exist. The first major problem to be solved is the need to avoid polarization fading. When the state of the polarization fluctuates in the fibre, the light power received at the detector changes, and this can result in significant reduction of the SNR—in the worst case, the signal can even totally disappear! A way to avoid this problem, using only a passive component, is to insert a fibre depolarizer at some point in (most conveniently at the end of) the sensing fibre coil. The depolarizer is fabricated simply using a short birefringent polarization-maintaining fibre, in which the two orthogonally polarized propagation modes have a difference in velocity. When the usual broadband LED (or a superluminescent fibre) wide spectrum (low coherence) light source is used to excite the FOG, the variations in differential mode delay result in a different output polarization state from the polarization-maintaining fibre section for each wavelength component. This effectively reduces the degree of polarization by 'scrambling' it, to give a different polarization at each wavelength, and to render it effectively unpolarized when the effect is averaged over the full bandwidth of the source (even though each individual wavelength component is still strongly polarized). Hence, using this 'depolarizer', the polarization fading problem can be overcome, as some wavelengths will still interfere without fading. Unfortunately, an undesirable polarization component is induced, which is perpendicular to the polarization axis of the polarizer. Due to the finite extinction ratio of practical polarizer elements, this configuration is not suitable for realizing FOGs of very high resolution. This configuration is therefore only used for low-cost, moderate-grade gyros.

To increase the optical efficiency, and hence the sensitivity, a polarization-maintaining fibre coil is usually used. With such a coil, undesirable coupling to the orthogonal polarization component is greatly reduced. Moreover, even if a small coupling were to occur, the two polarization modes in the fibre have different propagation velocities, so the undesired component has a large optical delay compared with the desired component. Therefore, when using a low coherent source, such as an ELED, the undesirable component cannot interfere with the desired one. This configuration is used for intermediate grade gyros. To improve the sensitivity more, a $LiNbO_3$ integrated optical circuit modulator can be introduced as shown in figure C3.1.19.

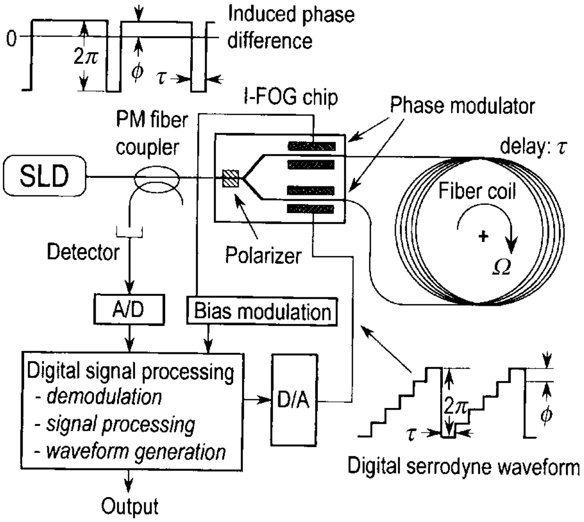


Figure C3.1.19. Schematic of a more sophisticated gyroscope arrangement with digital serrodyne modulation.

This integrated optical circuit is called the FOG chip, in which one coupling branch, two phase modulators, and a polarizer are all integrated together on a common planar substrate. This configuration is used for high-grade FOGs. When a proton-exchanged $LiNbO_3$ planar waveguide is used, it also acts as a polarizer with quite a high (>60 dB) extinction ratio. In this system of figure C3.1.19, a sophisticated signal-processing scheme has also been included, which is explained below.

A very small Faraday rotation effect can occur in the fibre-sensing coil, due to the Earth's magnetic field, and this results in another error factor in the FOG. The Earth's magnetic field lines are essentially parallel over the small dimensions of the coil, so the Faraday effect induced in one half of the coil would ideally be cancelled by that induced in the other half. However, when birefringence exists in the fibre coil due, for example, to bending or other mechanical stress, the FOG can suffer an error or drift. One way to reduce it is again to use polarization-maintaining fibre for the sensing coil. A polarization-maintaining fibre with high birefringence can effectively reduce the drift, because it prevents the lightwave suffering polarization rotation due to the Faraday effect. Another way to reduce the drift is to place a depolarizer at each end of an ordinary single-mode fibre coil. Recently, a general formula to describe the Faraday effect induced drift has been derived, in which the mechanism of the reduction of this drift has been theoretically derived.

If the temperature distribution in the coil changes with time, in a manner asymmetric with respect to the centre of the fibre length, the CW and the CCW lightwave will experience a very slightly different line-integral of temperature along the coil. Transient thermal changes of very small magnitude can give rise to a significantly different phase change in each direction in the coil, resulting in output drift. A temperature rate of change of only $0.01°\mathrm{C\,s^{-1}}$ can induce a very substantial drift of the order of $10\,\mathrm{deg\,h^{-1}}$ in an FOG with a badly designed coil! Special coil winding technologies have been invented, in which every section of fibre in the coil lies adjacent, and closely thermally coupled to a similar section of coil that is located at a symmetrical position with respect to the fibre centre (i.e. both sections lie at equal in-fibre distance from, but at opposite sides of, the coil centre). This makes the overall temperature distribution symmetrical with respect to the centre. To make a high-grade model, with $0.01\,\mathrm{deg\,hour^{-1}}$ sensitivity, such techniques have been applied.

For high-grade applications, such as aircraft navigation, a wide dynamic range of about 7 decades, and a good linearity of about 10 ppm are required. To realize these, a special signal-processing scheme, involving closed-loop operation, has been proposed and used.

The schematic configuration of this has already been shown in figure C3.1.19, and we shall now describe the signal-processing scheme. The phase modulation waveform shown in the figure is used to intentionally give a more complex phase-difference modulation between the CW and the CCW lightwaves. This phase difference occurs because of the different effective position of the modulator in each path and the phase modulation waveform used is called the *digital serrodyne waveform*. The modulation waveform can be thought of as having a 'saw-tooth' envelope, but with steps, as can conveniently be generated by a D/A converter. Each phase step is adjusted to correspond to one round-trip travelling time, τ, of the lightwave in the fibre coil. The amplitude is kept at 2π. After the CW and the CCW waves have both suffered the same time delay of τ, these two waves arrive with a phase difference at the detector, this difference being equal to the phase step ϕ in the serrodyne waveform. The Sagnac phase is then compensated, using this phase difference as input to a feedback control loop. Under the condition where the amplitude of the waveform is set to 2π, the phase difference ϕ is proportional to the frequency of the serrodyne waveform. Consequently, the input rotation rate is converted to a modulation frequency, which can then be measured with a frequency counter, to provide a wide dynamic range output from the sensor, which also has good linearity.

Applications of such gyros are expanding rapidly in various fields. The required sensitivity and dynamic range can vary greatly according to the application. FOGs have already been developed, in moderate, intermediate and high-grade forms, with progressive increases in cost and complexity. For moderate-grade applications, the analogue output of the detector is directly measured, this is called *open-loop operation*. For high-grade applications, the *closed-loop operation* method should be used.

Japanese gyro makers have created new application fields for industrial and consumer applications, for example, car navigation systems, and control systems for cleaning-robots, forklifts, agricultural machines and unmanned dump trucks suitable for hazardous environments. A camera stabilizer, to provide stable TV pictures from a helicopter, has also been developed with FOG technology for the sensor. Radio controlled helicopters with I-FOGs have also been produced for agricultural applications, such as planting seeds and spraying chemicals. A North-finder 'optical fibre compass', using an open-loop FOG has been developed. The National Aerospace Development Agency of Japan used an inertial sensor package with FOGs, for the first time, in their rockets, for micro-gravity-mission experiments. The FOG was selected for this mission as a silent gyro. The first launching took place in 1991, and was the first application of the FOG in space. The Institute of Space and Astronautical Science, Japan has developed a rocket having an inertial navigation system (INS) with closed loop FOGs. The first successful flight of this M-V rocket was on 12 February 1997, using an additional radio-wave guidance technique. The M-V-1 launched a Satellite MUSES-B, with a mission to construct a VLBI (very long base-line interferometer) for radio astronomy, when in radio contact with other antennas on the earth. Also, in this satellite, an open-loop FOG, having a $0.05\,°\,h^{-1}$ bias stability, was used for rate control.

It should be pointed out that the Boeing 777 uses a inertial navigation system which combines the use of six more conventional ring-laser gyros, of $0.01\,°\,h^{-1}$ grade, with four more recently developed all-polarization-maintaining-fibre open-loop FOGs, having $0.5\,°\,h^{-1}$ capability.

Applications requiring a sensitivity even greater than $0.001\,°\,h^{-1}$ exist, such as for space applications and ship navigation. Potential applications include deep-space and precision spacecraft navigation, and space pointing, and stabilization. However, for such higher-grade applications, the ELED that is commonly used to excite simpler FOGs has insufficient power and lacks wavelength stability. Because of this, laser-pumped superluminescent Er-doped fibre sources have been developed, and, using such a source, higher power and extremely high wavelength stability of typically a few ppm $°C^{-1}$ can be obtained. A rotation resolution better than $0.001\,°\,h^{-1}$ has already been demonstrated by several companies, but requires carefully temperature-stabilized conditions.

Fibre optic hydrophones

We shall now discuss fibre sensors for detection of acoustic signals in water, as required for many military (detection of marine vehicles) and civil applications (seismic oil exploration).

When an acoustic wave impinges on a fibre in water, the sound pressure induces change in its density and length, which results in a phase change of the lightwave propagating in the fibre. The change is of course very small in quiet seas, but it can be detected using interferometric configurations with a very long length (100–400 m) of sensing fibre. Using the compensation schemes for slow drift of temperature and mechanical strain, which were described above, a highly sensitive acoustic wave sensor in water, called a hydrophone, has been developed. As we discussed above, signals, as small as 10^{-6} radians or less, can be detected, representing an optical path length change of only 10^{-12} m at 1 μm. This corresponds to a change in the physical length of the sensing fibre which is only a small fraction of the diameter of a hydrogen atom! It is clear this will give tremendous acoustic sensitivity. Best sensitivity is achieved using mechanical amplifiers, such as mechanically compliant (e.g. made of easily deformable, perhaps even air-filled, materials) rods as coil formers. The fibre is wound around these, such that when acoustic waves interact, a compressive change in diameter of the soft rod is transferred to length changes in the fibre, giving a much greater phase change for the same acoustic influence, than would occur for bare fibre.

By using a wide variety of sensor multiplexing techniques, such as time-division multiplexing (TDM), large arrays of such hydrophones have also been realized. These will be discussed in more detail later in sections C3.1.4 and C3.1.5 of this chapter, and the textbooks and the review paper by Kersey in the bibliography adds yet more detail.

The fibre optic current sensor

High-voltage power systems often deal with enormous voltages (as high as 750 kV or more!) and very large currents, often several kA. In this area, optical fibre current measurement schemes [148] can provide excellent electrical insulation and almost total immunity to EMI. Optical fibre current sensors (OFCS) satisfying such requirements are soon expected to take the place of traditional electrical current transformers (CT). After many years of careful research, rigorous field tests of the OFCS are finally showing good performance, and the research now seems to be in the final engineering stages.

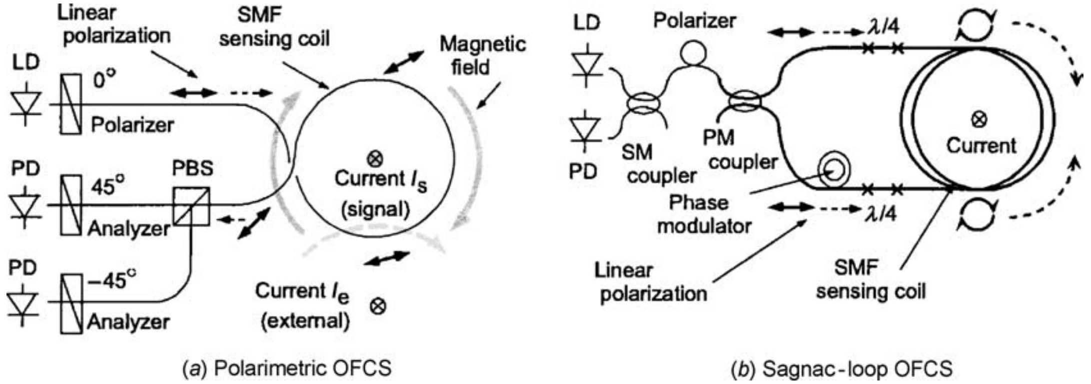
The OFCS measures current indirectly, by measuring the rotation of the state of optical polarization (SOP) induced by it by the total magnetic field component along the axis of the sensing fibre. Figure C3.1.20(*a*) shows a typical configuration of a polarimetric OFCS. The sensing fibre coil is wound around a current-carrying conductor, and linearly polarized light is launched into the coil. The SOP rotation is due to the Faraday effect of the magnetic field, in the direction of the fibre axis, the magnetic field being induced by the current-carrying conductor inside the fibre coil. The plane of polarization of the propagating HE_{11} mode in the fibre is rotated through an angle, ϕ, given by:

$$\phi = V\int H\,\mathrm{d}l$$

where ϕ, V, H and $\mathrm{d}l$ are, respectively, the Faraday rotation angle, the Verdet constant, the axial magnetic field component and the length along the fibre. Because the sensing fibre coil is formed into closed path, Ampere's law gives the relation:

$$I_\mathrm{S} = \oint H\,\mathrm{d}l$$

where I_S is the current in the electrical conductor passing through the fibre sensing coil. Suppose the fibre has a number of complete turns, n, then, from the two equations above, the relation:



(*a*) Polarimetric OFCS (*b*) Sagnac-loop OFCS

Figure C3.1.20. Two types of optical fibre current sensors: (*a*) using polarimetric interferometer and (*b*) using a Sagnac interferometer.

$$\phi = nVI_{\mathrm{S}}$$

is obtained. These form the well-known fundamental equations for the OFCS. These relationships are, however, only valid when the sensing fibre coil has no linear birefringence. If significant linear birefringence is present, the polarization state in the fibre is disturbed, and the output of the OFCS is changed, giving measurement errors. To overcome such problems, and reduce the linear birefringence of the fibre, several solutions have already been developed. Twisting a single mode fibre during the winding process is one method, and thermally annealing the fibre coil, to relieve strain induced by the winding process, is another. Additionally, a low birefringence fibre using flint glass has also been developed, which also has a stronger Verdet constant. This has been demonstrated to give a high performance, suitable for meeting practical engineering requirements [52].

The configuration shown in figure C3.1.20(*a*) is only a polarimetric interferometer, i.e. mixing of polarization components is enabled in a polarization analyser. This should ideally have its axis inclined at 45° to the axis of each interfering beam. Another configuration for the current detection has been developed, in which a Sagnac interferometer arrangement, similar to that used for the fibre optic gyro, is adopted. The configuration is schematically shown in figure C3.1.20(*b*). In this system, the SOP in the sensing fibre coil is arranged to be circular by using quarter wave elements at both ends of the coil. In this configuration, the circular SOP is maintained throughout the propagation in the coil, but, via the Faraday effect, a phase difference is induced between the CW and the CCW travelling in the fibre coil. The phase difference is read out from the interference signal at the detector, using the same type of signal processing as the FOG. Hence, all the sophisticated and compact optical and signal processing modules, already developed for use with the gyro can also be used in this version of the current sensor.

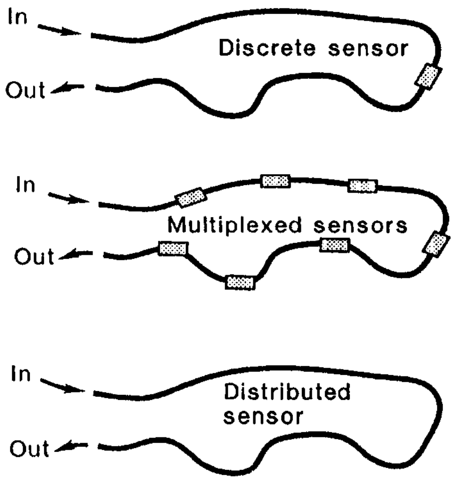
We shall now conclude our discussions on the different types of single-point (discrete) sensors. The above discussions have made occasional reference to multiplexed or distributed sensors, but the following two sections will now concentrate on these aspects more fully.

C3.1.4 Multiplexed optical fibre sensor systems

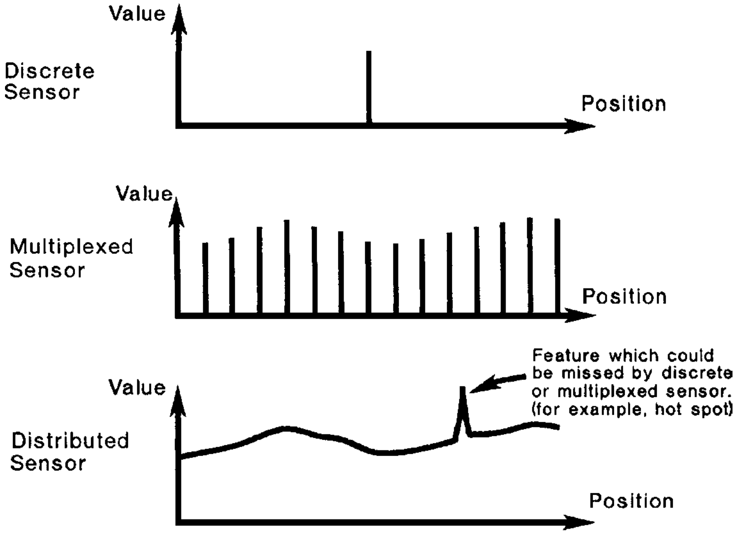
C3.1.4.1 Introduction

This section shows how sensors may be multiplexed [153] and the following one will deal with truly distributed [154] optical fibre sensors. It is convenient to define multiplexed sensors as those designed to collect data from a number of discrete sensing points, or sensing regions, and distributed sensor types as those that operate on a continuous length of fibre, and are capable of determining the variations of a desired parameter along the length of the fibre as a continuous function of distance. Figure C3.1.21 depicts the basic concepts of distributed and multiplexed sensors. The upper curve, figure C3.1.21(*a*), shows a schematic of an in-line example of both types (fan-out topologies and many more arrangements are possible for multiplexed sensors) and the lower curve, figure C3.1.21(*b*), shows the type of response to monitor the values of parameters, that each type can provide, as a function of physical location.

Extending the capability of a single measurement terminal, to address a multiplexed array of many passive sensing heads, rather than a single one, not only makes the use of a more complex monitoring station more economically viable, but it can also lead to a more accurate and reliable comparison (see figure C3.1.21(*b*)) of values of the measured parameter, because the same interrogator is used to read each one. The distributed sensor (bottom curve, figure C3.1.21(*b*)) however, goes a stage further, allowing the full distribution of the measured quantity to be determined with no gaps in coverage. In both multiplexed and distributed cases, one or more passive sensors, or sensing-fibre sections, with controlled environmental parameters, can be used as a calibration aid for the interrogator.


In
Out
Discrete sensor
In
Out
Multiplexed sensors
In
Out
Distributed sensor

(*a*) Schematic illustration of discrete, multiplexed and distributed sensors


Value
Discrete Sensor
Position
Value
Multiplexed Sensor
Position
Value
Distributed Sensor
Position

Feature which could be missed by discrete or multiplexed sensor. (for example, hot spot)

(*b*) Advantages of distributed sensors

Figure C3.1.21. Schematic illustration showing (*a*) an in-line configuration of multiplexed and distributed sensors and (*b*) the different types of spatial coverage offered by each type of sensor.

As with multiplexing in communications, the ultimate information gathering capacity is fundamentally limited by the available bandwidth and by the signal/noise ratio of the detected optical signals, and similar care must be taken to avoid undesirable crosstalk between signals from apparently independent information sources; represented in this case by the individual sensor elements.

In describing multiplexed sensors, we might perhaps start with a historical note. One of the earliest schemes for sensor multiplexing in optical fibres was reported by Nelson *et al* [53], where a fibre optic branch-tapped network was used, in conjunction with an optical time-domain reflectometer (OTDR). This system, which will be described later, is capable of receiving and independently monitoring the separate returns from a series of reflective sensors.

The following sections will describe many more methods, outlining the most significant developments in the technology, with the sensors being grouped or classified according to the method used to address the various sensing elements.

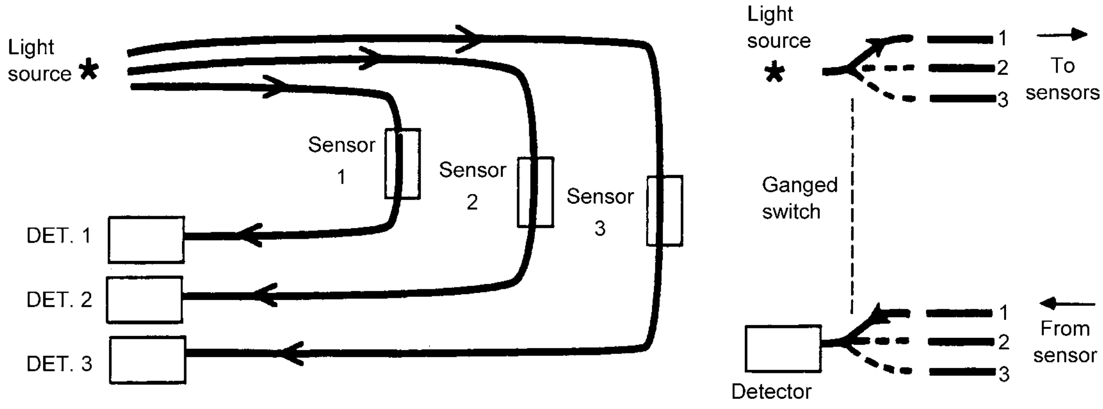
C3.1.4.2 Spatial multiplexing (separate fibre paths)

Starting with the simplest low technology approach, the technique of using separate fibres to communicate with each separate sensing element (figure C3.1.22), although trivial in scientific terms, is virtually guaranteed to avoid one of the pitfalls of multiplexing, that of crosstalk between sensors.

As it is the easiest method to implement, it was one of the first to have been used for practical applications. Perhaps the most unusual and dramatic use so far has been the application of 152 separate graded index fibres, each 0.6 km long, for nuclear weapon diagnostics [54]. For this particular application, light was generated at the sensor head end, so the light source and outgoing fibre paths shown in figure C3.1.22 were not required, and the multiple fibre channels were used to guide light from the event. Each individual receiving fibre guided light to produce a 'pixel' of an image on a fluorescent phosphor imaging screen, presumably located in a monitoring area at a safe distance!

For cost-sensitive applications, where a slow update is acceptable, it is convenient to incorporate a fibre switch into a single optoelectronic processor unit, with a single light source and detector to permit it to 'poll', i.e. sequentially connect to, each of the sensing heads in turn (figure C3.1.22(*b*)).

A multiplexed system for monitoring fluorescent dye concentration in water, via separate 600 m lengths of optical fibre to each sensing head, has been reported [55]. In this system, a mechanically scanned mirror system was used to inject light from an exciting laser sequentially into each separate sensor head via separate fibres, the returning fluorescent light travelling through separate fibres to a common photomultiplier detector. This system was capable of monitoring fluorescent dye concentration down to levels of 10^{-10} by weight in water.



(*a*) Basic arrangement (*b*) Switched arrangement

Figure C3.1.22. Spatial multiplexing using separate fibre paths: with (*a*) fixed paths, (*b*) with fibre switches to enable one source and detector to be used.

Although usually used to address independent sensors, multiple fibres have been used to permit precision digital sensing in systems where each 'bit' of binary-coded information is carried by a separate fibre. This technique has been used in simple position-encoding schemes, using binary-coded discs of the type used in conventional optical shaft-angle encoders [56]. A more complex system has been described for the 'digital' measurement of temperature, using a series of isothermal (i.e. all in close thermal contact) birefringent crystals, each sensing crystal having a different length according to the significance of the desired order of the 'bit' of binary-coded information [57]. The temperature-dependent birefringence of each is separately monitored via a polarization analyser, which converts these changes in the birefringence into changes in the amplitude of the transmitted light, each with a different sinusoidal response to temperature. The separate outputs were combined to produce the equivalent of a binary word, which defined the temperature of the crystals.

C3.1.4.3 Time-division multiplexing (TDM)

The concept of TDM, in combination with high-speed digital technology, has revolutionized modern-day communications. It is hardly surprising that it is also an attractive choice for multiplexing sensors! The time differences between signal returns from each sensor, necessary for a TDM system, are usually achieved by arranging differing total optical propagation delays for the signals from each sensing element, using extra coils of fibre, whenever longer delays are needed. In addition, in order to distinguish, and separate, the signals from each sensor, it is necessary to modulate the light source with a temporally varying signal.

The simplest form of encoding, for ease of both modulation and demodulation, is to use a repetitive pulse, having a short duration so that the set of returning differentially-delayed pulses from each sensor does not coincide at the detector. Also, it must have a pulse repetition rate that is low enough to allow each reflected pulse 'echo' to return back from the most distant element of the sensing array before the first of the subsequent set of reflected pulses returns from the nearest sensing element.

One of the first multiplexing methods suggested for use with hydrophone sensor arrays was of this type [58]. One of their proposed methods is shown in figure C3.1.23(*a*). This used two parallel fibre-optic, cross-coupled, highways, the first to distribute a transmitted optical pulse and tap a portion of it into each transmissive (loss-modulation type) sensor, and a second highway to collect signals from each individual sensor and guide the set of returning pulses to the detector.

The second of their proposed arrangements used a single tapped highway, with reflective sensors, each having a measurand-dependent reflection, connected to this highway with directional fibre couplers (figure C3.1.23(*b*)). This array was interrogated using a conventional OTDR arrangement, with a semiconductor laser source and an avalanche photodiode detector. The basic method of optical time-domain reflectometry, which we mentioned briefly in the chapter introduction, was devised by Barnoski and Jensen [59] and researched further by Personick [60]. The concept is depicted in figure C3.1.24, in a configuration commonly used to monitor losses in optical fibres and reflections from discontinuities. It is based on an optical radar (or LIDAR) concept, where a short pulse of light is launched into a fibre waveguide and variations of backscattered light signal with time are monitored.

Usually, light from a pulsed semiconductor laser (or Q-switched fibre laser light source) is launched into a section of fibre via a directional coupler, which serves also to direct a portion of the backscattered light fraction, returning from the fibre on test, to a high-speed PIN-FET or avalanche photodiode (APD) detector. The time of flight determines the distance, and the intensity variation normally indicates properties of the fibre under test (see lower curve of figure C3.1.24). It has become a standard test-gear

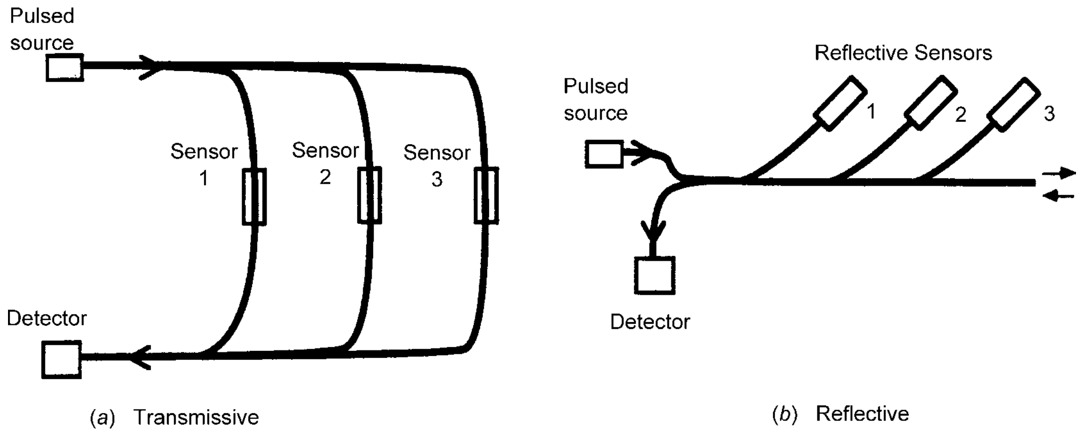


Figure C3.1.23. TDM of fibre sensors: (*a*) ladder network with transmissive sensors and (*b*) branched network with reflective sensors.

instrument for use by optical fibre engineers and researchers, to examine the continuity and attenuation of optical fibres, and observe reflections from fibre breaks, connectors or other discontinuities.

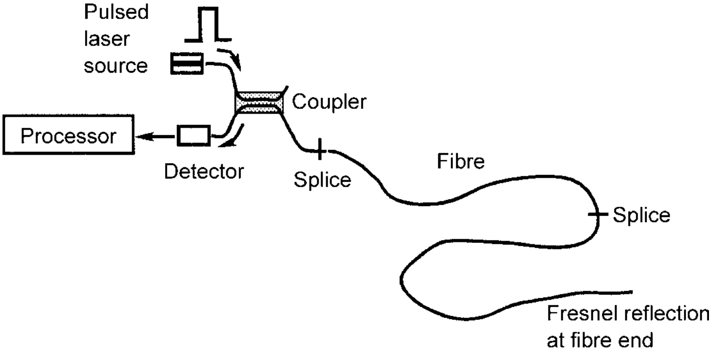
If used to interrogate multiplexed sensors, it is possible to measure variations in either the reflected power from, or the transmission loss in, each sensor element (distributed sensing will be discussed in more detail later). Desforges *et al* [61] have also reported experimental results with reflective sensors using an OTDR, but in their case the sensors were located at regions where the optical fibres are deliberately bent, to cause non-invasive coupling of light into and out of a continuous fibre to the reflective sensors.

The OTDR may also be used to monitor discrete loss-modulation sensors, of the 'microbend' type we discussed earlier, situated along the length of a continuous optical fibre. This technique has been proposed to monitor strain or deformation in civil engineering structures [62, 63, 148] and for non-destructive testing of engineering composite materials [64]. However, care must be taken with such multiplexed arrangements if quantitative results are required, because such microbend sensors cause severe mode conversion in multimode fibres, and the loss of each such sensor is strongly dependent on the mode excitation of the fibre immediately before the micro-bent section. Thus, when multimode fibre is used to interconnect and form an array of closely located sensors, there will, in general, be significant inter-element crosstalk in the response.

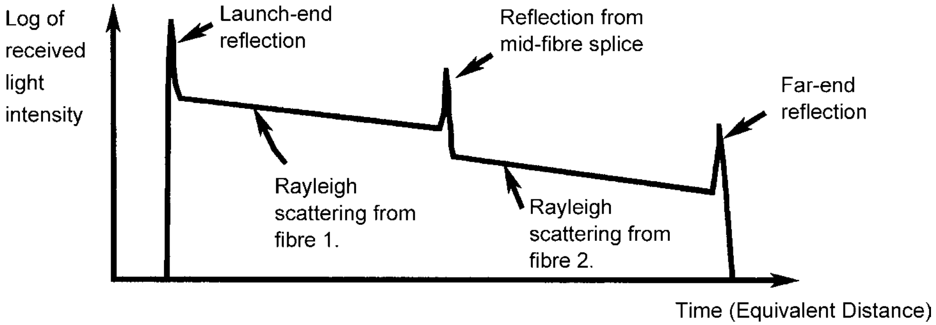
The main disadvantage of a simple pulsed source for addressing multiplexed arrays is the poor optical efficiency due to the low excitation duty cycle. This problem is exacerbated, either by the peak-power limitations of semiconductor light sources, or by the onset of non-linear optical processes in monomode fibres. A similar problem has, however, already been previously met, and dealt with, in radar systems, where it was also desired to distinguish differentially delayed signal returns without need for enormous transmitted power levels[†]. The radar technique most relevant to time-division optical sensor multiplexing is that where the transmitted radio signal has a pulse-code intensity modulation envelope

---

[†] It should be noted here that the radar field [65] is one from which numerous techniques, applicable to multiplexed and distributed optical sensors, have been drawn. This source of inspiration for optical sensor ideas will be referred to again later.



(*a*) Basic arrangement of optical-time-domain reflectometer (OTDR) system



(*b*) Intensity versus time, OTDR return

Figure C3.1.24. Basic concept of the optical time domain reflectometer, or OTDR, showing (*a*) optical arrangement and (*b*) typical OTDR returns from a fibre.

that has a 50% duty cycle and has an autocorrelation function corresponding to that of a periodic pulse train [65]. Simple cross-correlation of the detected signal, with differently phased versions of the code originally transmitted, can then be used to separate the pulses corresponding to the individual sensors, just as if a single pulse of much higher peak power were to be used.

An alternative means of reducing the peak power requirements for the optical source (again one commonly used in radar systems) is to produce a temporally extended ‘chirp’ signal from an electrical pulse, using a dispersive surface acoustic wave (SAW) filter, and to use this signal to modulate the intensity (amplitude) of the light source [58]. The detected return signal is then subjected to an inverse transformation, using a second SAW filter, thereby reconstituting a replica of the originally transmitted pulse from each reflected signal; the net result from a sensor array, with delays of different lengths, being a time-division multiplexed stream of pulses, similar to those which would be obtained without the SAW filters.

As mentioned earlier, the TDM technique is also attractive when used with interferometric sensors. If an optical heterodyne method is used, where the returning signal is mixed with a reference signal of shifted frequency, the amplitude of the returning signal will, provided it remains constant, have no effect on the phase of the resulting beat signal. Such sensors, therefore, will require no amplitude referencing, as the output of the sensor is represented by the phase of the beat signal. A multiplexing scheme of this nature was first reported for hydrophone applications by Dakin *et al* [66]. This particular implementation (figure C3.1.25(*a*)) involved the launching of a consecutive pair of optical pulses, each having slightly different frequency, into a linear array of interferometric intrinsic fibre strain sensors, each joined by partially reflective splices. The intrinsic sensors are coils of monomode fibre, usually potted into polymer cylinders, to enhance their response to acoustic pressure waves. The initial optical pulses transmitted were consecutive, i.e. did not overlap in time, but the differential delay of returning pulses from adjacent splices gave rise to coincidence of the second pulse, from the nearer splice of a sensor element, with the first pulse, from the subsequent splice (figure C3.1.25(*b*)).

Thus, with suitable pulse separation and duration, a time-division multiplexed stream of heterodyne beat signals was obtained from the receiving detector, each heterodyne pulse corresponding to an element of the array, and carrying phase modulation proportional to the changes in optical path length. These path length changes were a direct measure of the strains arising from insonification of the corresponding fibre hydrophone element. Phase de-modulation of each time-demultiplexed channel yielded the acoustic signals, free from any dependence on the amplitudes of the reflected light signals. In a patented modification, the same research group [67] first showed that it is possible to compensate the unbalanced interferometer arrangement using an optical loop of fibre as a ‘pre-delay’, thus balancing optical paths and greatly reducing the effects of undesirable phase noise that arise from frequency fluctuations of the source laser (see later section C3.1.4.7 on frequency-modulated carrier wave (FMCW) methods).

The attraction of such coherent heterodyne TDM approaches is the great sensitivity to even tiny phase changes that can be achieved using the coherent detection process and, secondly, the excellent dynamic range that is possible using electronic phase demodulation of the intermediate frequency signal. The improvement in the detection process is greatest if a returning reflected signal is mixed with a strong local oscillator signal, derived at the monitoring station from the initial source [68]. The advantage of *differential* sensing of the distance between adjacent reflective splices, achieved by the method shown in figure C3.1.25, may theoretically be retained if a three-wave mixing process is performed [69]. Using this latter method, the beat signal between two weak received signals should be recoverable, after mixing with a strong local oscillator signal on the detector.

Based on many of the concepts arising from those early publications, the field of marine acoustic sensing has been one of the major practical success stories in the field of optical fibre sensors. The applications are in two main areas, firstly for naval surveillance applications (all the usual ones of towed arrays, vehicle arrays and fixed sea-bed arrays) and secondly for seismic surveys, where intense sound sources, often explosive ones, are used to investigate sub-sea rock strata for oil-bearing features. It now appears very likely that optical fibre hydrophone sensor arrays, having a conveniently passive all-fibre ‘wet-end’, will completely take over from older technology using piezo-electric sensors. The latter require electrical pre-amplifiers, complex electrical wiring to electronic multiplexers, and a sophisticated electronics communications system. Electrical systems are difficult to design and reliably maintain in a corrosive and conductive salt-water environment. An excellent review of the subject has been given by Kersey in bibliography 1, vol 4, chapter 15. An advanced seabed sensor array system and a working sea test are described in papers by Nash *et al* [70] and Cranch *et al* [71].

C3.1.4.4 Wavelength-division multiplexing (WDM)

The use of WDM, unlike the alternatives of TDM and sub-carrier frequency-division multiplexing (FDM) techniques, has the advantage that there is no *theoretical* loss penalty when compared with the single-fibre-per-sensor approach. The method involves guiding optical power to each sensor, and back to a corresponding sensor, via a route dependent on the wavelength designated for the interrogation of

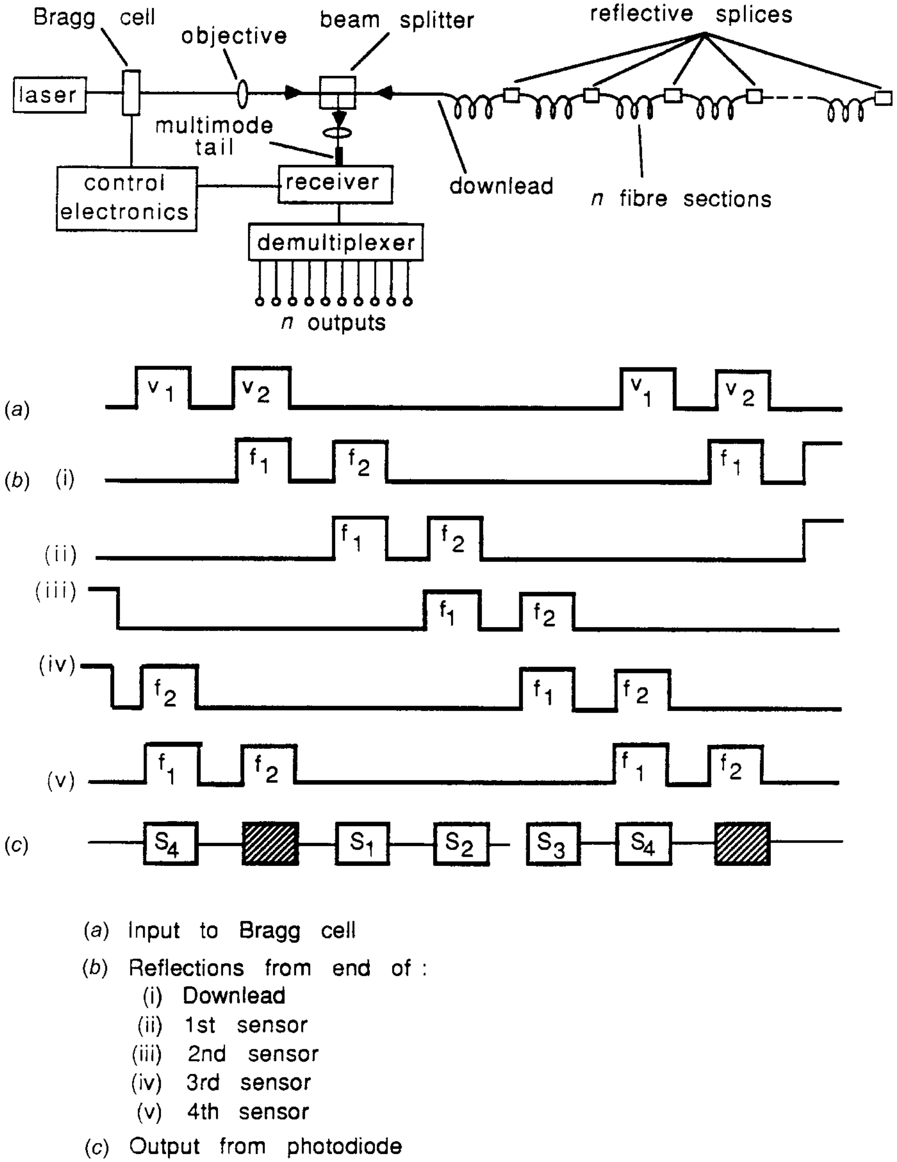

Bragg cell
objective
beam splitter
reflective splices
laser
multimode tail
control electronics
receiver
downlead
n fibre sections
demultiplexer
n outputs
(a)
v1
v2
(b) (i)
f1
f2
(ii)
(iii)
(iv)
(v)
(c)
S4
S1
S2
S3
(a) Input to Bragg cell
(b) Reflections from end of :
(i) Downlead
(ii) 1st sensor
(iii) 2nd sensor
(iv) 3rd sensor
(v) 4th sensor
(c) Output from photodiode

Figure C3.1.25. (*a*) Diagram of optical fibre hydrophone array. (*b*) Timing diagram for a four-element array with: (1) the input to the Bragg cell; (2) the reflection from the end of (i) the downlead, (ii) the first sensor, (iii) the second sensor, (iv) the third sensor and (v) the fourth sensor and (3) the output from the photodiode.

that particular channel (see figure C3.1.26 for a very basic schematic). The path of light to each sensor is directed using WDM coupling components, similar to those used in communications systems. These components are in theory lossless, but will, in general, introduce a small loss in practice (typically 1–3 dB for each pass), imposing, therefore, a very small penalty in an otherwise lossless multiplexing method.

Although the spectral width of the fibre transmission ‘window’ is potentially enormous, when compared with the very much lower information rate theoretically required for telemetry, the potential

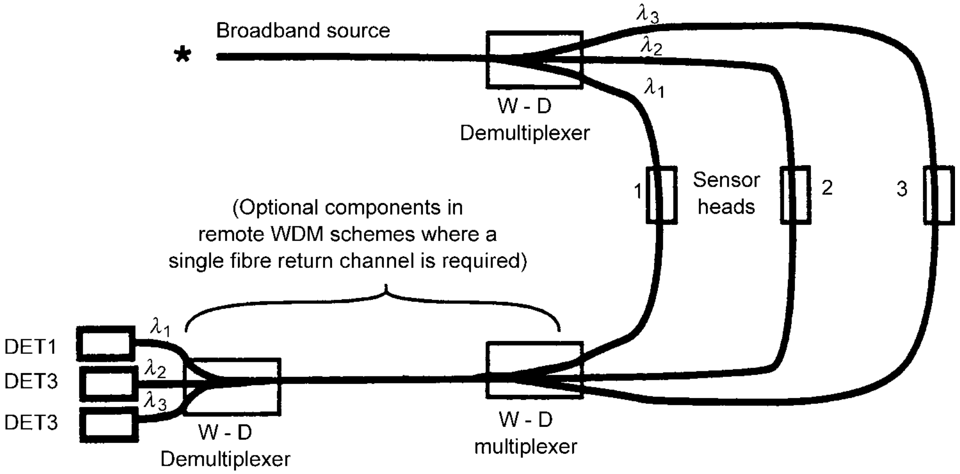


Figure C3.1.26. Schematic diagram of a wavelength-division multiplexed arrangement for remote detection of the state of three amplitude-modulation sensor heads.

of multi-channel WDM was originally rather limited by practical difficulties in achieving sufficient selectivity in the WDM filters. This has changed dramatically with the excellent wavelength routing components that have now been developed for multi-channel WDM telecommunications systems. Using broadband LED or superluminescent fibre sources, or multi-channel laser sources, the practical number of independent channels in the fibre window is now potentially very large. It is possible to use 'spectral slicing' techniques with either LED [72] or superluminescent fibre sources [73]. This latter technique involves filtering out two or more separate narrow channels, from within the broad spectral linewidth of the LED (in this case for distributed sensing, but the basic concept is the same). The separation can be performed using narrow-band filters—a method that inevitably decreases the average signal in each channel by a factor at least in proportion to the number of channels required.

Another method of achieving a large number of channels is to use a number of independent narrow-linewidth laser sources, which removes this inherent loss penalty, as modern WDM combiners can, combine channels with zero loss in theory, and in practice they have close to 100% efficiency. Generally, the limit with lasers is set by practical considerations, such as the selectivity of channel filters, the availability, selection and stabilization of laser sources, and the cost of a multi-laser system of this nature. Again very dramatic developments in the light-sources (e.g. DFB fibre lasers) passive components (WDM splitters, add-drop multiplexers, etc) and many other sub-systems have been made to facilitate use of WDM in telecommunications applications, which paves the way for such methods in sensors.

The WDM approach was first used to permit the addressing of separate 'bits' in a ten-bit digital sensor, designed to measure a *single* physical parameter, in this case the angular position of a shaft in a fibre optic shaft angle encoder [74]. A broadband white light source in the transmitter/receiver terminal launched light into the outgoing fibre to the sensor head (shaft angle encoder), and two diffraction gratings were used, the first one in the encoder, firstly to separate the white light from the outgoing fibre lead into ten wavelength channels, and then to recombine the reflected code-carrying channels after intensity modulation of each 'bit' by the coded disc. A second similar diffraction grating was used again in the decoder, to direct each of the returning signals, according to their wavelength, onto a separate detector of an array. The digital encoder disc was a 'Gray-coded' disc, as commonly used in conventional optical encoders, but in this version a different colour was used to interrogate each separate coded band on the encoder disc, hence allowing transmission through a single optical-fibre cable (or two-way send/return cable) by wavelength multiplexing.

The multi-channel WDM approach has also been proposed to provide additional picture elements for remote imaging [75]. In this case, a diffraction grating system was again used in the sensor head, to enable the spatial position of each pixel of a one-dimensional white-light image to be uniquely coded, according to its optical wavelength, before transmission to a remote monitor station. At this point, the image is recreated, using a second diffraction grating to perform the inverse transformation. This way, one dimension of the two-dimensional image is divided according to wavelength. The other dimension can be divided easily using a one-dimensional array of separate fibres.

Now that WDM systems have become so widely accepted for fibre-optic communications systems, the practical capability of WDM systems to address large arrays sensors is becoming ever greater and yet more economically viable. It is applicable to many types of optical sensor, and there are now many key components for WDM routing, including multi-line laser sources, and low-loss multi-channel filters, grating filters, add-drop filters, to name but a few. An example of a multi-element hydrophone array, using a combination of both TDM and WDM methods, has been presented by Vohra *et al* [76].

C3.1.4.5 Multiplexing of in-fibre Bragg gratings using TDM and WDM methods

In view of the importance of Bragg grating sensors, it is appropriate to discuss how these sensors can be multiplexed. In view of their narrow reflective spectrum, they readily lend themselves to WDM methods. Secondly, the ability to write them at any point in a fibre, with associated position-dependent variations in the two-way optical delay from a pulsed light source to the sensors, according to their position, allows TDM to also be used. Space here does not permit a full discussion of what is now a major research area, but the excellent review by Kersey *et al* in the bibliography covers this area very extensively.

Some of the methods used to address multiple gratings, all situated in a single fibre cable, are listed below, many of which were of course mentioned in the earlier section describing how individual gratings may be addressed:

- Use of broadband source launched into fibre, and a spectrometer (CCD spectrometer or Fourier-transform spectrometer) to interrogate the reflected spectrum (system as in figure C3.1.12, but with multiple in-line Bragg gratings in the fibre).
- Use of a scanned narrow-band optical filter (examples: conventional bulk-optic Fabry–Perot, all-fibre Fabry–Perot, acousto-optic tuneable filter, scanned Bragg grating) in conjunction with a broadband source launched into fibre, and a detector to receive reflected light from arrays of in-fibre Bragg gratings. The filter may either be scanned, to measure the peak reflectivity of each grating in turn, or, with the aid of a feedback loop, may be locked on to track each grating in turn.
- Use of a mechanically-scanned (fibre is stretched with a PZT) all-fibre Michelson interferometer, followed by a single optical detector, to interrogate reflected signals from arrays of in-fibre Bragg gratings. Fourier transformation of the detected output signal is performed, in the same manner as used in the well-known Fourier transform spectrophotometer using a bulk-optics interferometer [77].
- Rare-earth-doped fibre laser, operating multi-wavelength, using the Bragg gratings as end-mirrors.

Both the review by Kersey and another by Dakin and Volanthen [78] cover many of the pitfalls that have to be avoided to prevent undesirable measurement errors. Examples of potential problems include:

- The optical filtering effect that in-line fibre gratings have on light passing on to, and returning back from later sensors in the line. Spectral overlap, over the entire working range of the grating (which must include any wavelength shifting due to strain or thermal effects) has therefore to be avoided.
- Polarization effects, when gratings exhibit birefringence, and so have slightly different reflective spectra for each of the principal polarization directions. This causes problems if either the light source or spectral interrogator is polarization dependent.
- Spectral variations in the output of the source, in any of the passive coupling components, or in the final detector system. Any gradient in the curve of optical response against wavelength in the interrogation set will give rise to an undesirable offset error, i.e. a change in the apparent wavelength of the grating.
- Changes in the spectral shape of the grating, particularly spectral broadening, if it is not uniformly strained along its length.

C3.1.4.6 Sub-carrier FDM

In point sensors, which are connected by separate fibre paths to a common detector, the technique of FDM of the intensity modulation waveform can be used, modulating the electrical signal to drive a light source. Such a modulation may be used to facilitate separation of the outputs from individual channels in the detected signal. The simplest way of using the technique is the relatively trivial method of transmitting signals from separate light sources, each modulated by electrical signals of different frequency, via separate fibres to each of the sensor heads. The outputs may then be combined (or added) into a common output fibre and detected on a common detector, yet still be separable by frequency-selective electrical channel filtering. However, a more elegant FDM approach, using a single LED source, has been devised by workers at the University of Strathclyde in Glasgow [79]. This method, using a simple transmissive system, is shown in figure C3.1.27(*a*). The three sub-carrier modulation signals add at the detector output, as shown in figure C3.1.27(*b*), with the resultant of their vector addition being dependent on their relative phase angles. These phase angles are a function firstly of the original phases of the transmitted envelope modulation and secondly of the different delays they experience in transmission through the optical network containing the point sensors.

The term FDM is generally used in optical communications systems to imply systems in which an optical carrier is amplitude modulated by a composite electronic signal, consisting of a sum of modulated *sub-carrier* signals with different frequency channel allocations. In order to provide a more precise terminology, the prefix 'sub-carrier' has been inserted in the title of this section, but for the remainder of this review, the abbreviation FDM will, for convenience, be used to describe such systems.

These signals may be separated, firstly by employing multi-channel phase-sensitive detectors on the detected optical signal (using the original applied modulation signals as electronic reference signals), and then by solving simple simultaneous equations (linear regression) on the resulting scalar quantities. This process allows removal of any crosstalk terms from light that has passed through the sensor heads of each other channel.

The above system has the advantage over TDM methods of having 100% duty cycle and using much simpler and slower electronics. For low-frequency sensors, it should be easy to construct with a low-noise bandwidth in each of the phase-sensitive detection stages. However, although the electronics do not require a high-frequency response, care must be taken to ensure stability of electronic and optical delays, as serious crosstalk could otherwise result.

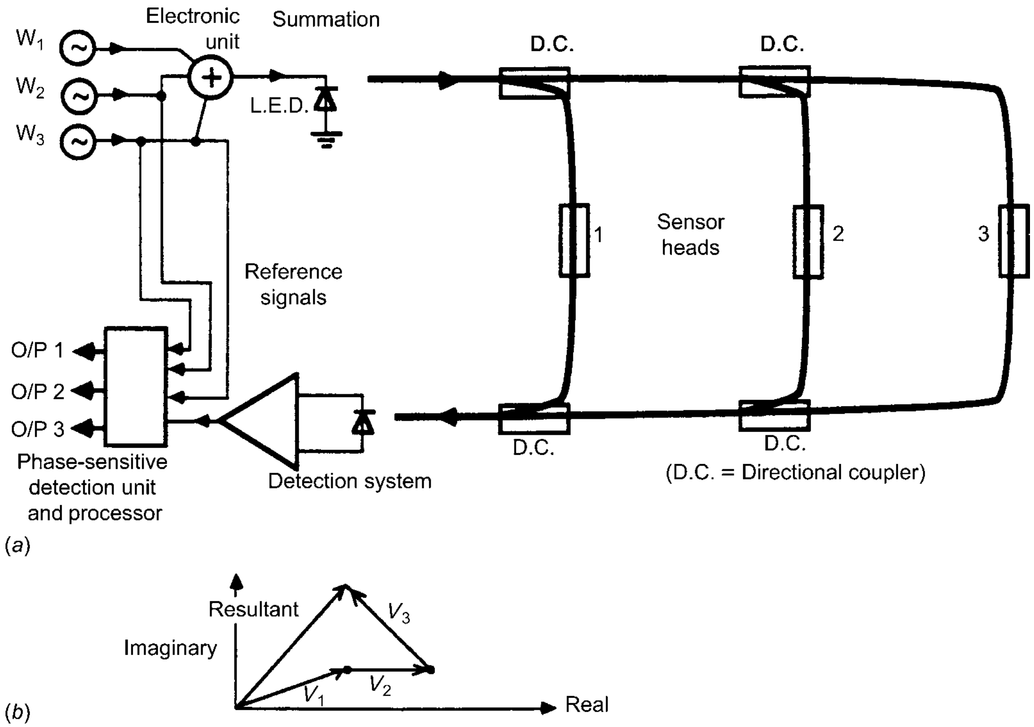
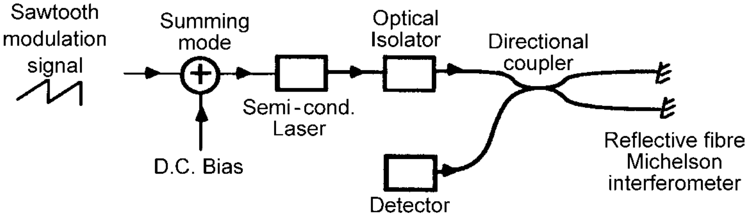


Figure C3.1.27. Multiplexing using frequency-division-multiplexing of sub-carriers (intensity modulation signals), showing: (*a*) schematic diagram of arrangement for three amplitude-modulation sensor heads and (*b*) vector diagram for the phase-sensitive summation of sub-carrier signals on an optical detector.

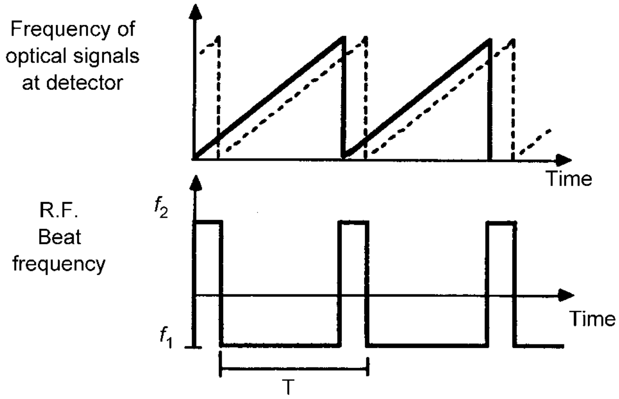
C3.1.4.7 The FMCW multiplexing scheme

This multiplexing method, like many of the preceding techniques, was adapted from the radar field and has much in common with the FDM method discussed in the previous section. However, in this case, the RF sub-carriers signals only actually appear at the optical detector, as a result of ‘beating’ or ‘mixing’ of optical waves, where difference frequency (heterodyne) signals are generated due to the ‘square-law’ characteristics of optical detection. The transmitted signal in the FMCW method is an optical carrier wave, the frequency of which increases (or decreases) linearly for a period, *T*, after which time it flies back to its initial frequency, before repeating the process. When now the source is connected to an interferometer with a differential path delay, the return signals on the detector differ by a frequency proportional to firstly the optical path difference they have experienced and secondly the frequency slew rate of the source (see figure C3.1.28).

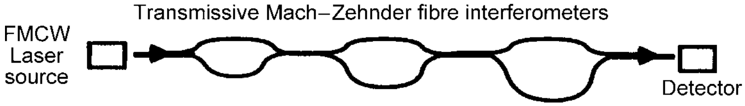
Such a system was first proposed in single sensor form by Uttam and Culshaw [80], with a later description from Giles *et al* [81]. If the source is connected via a series of interferometers, each having different delay paths [82], a series of RF carrier signals are produced at the detector by the heterodyne mixing process. The various sensor output signals may be isolated by electronically filtering out the corresponding frequencies. The sensor output signals may be represented by the phase, frequency or amplitude of the recovered RF signals, depending on the scale of the path length change (small path length changes are more readily discernable as phase changes, large path length changes more conveniently as frequency changes) or the transmission changes occurring in the sensor.


Sawtooth modulation signal
Summing mode
Optical Isolator
Directional coupler
Semi-cond. Laser
D.C. Bias
Detector
Reflective fibre Michelson interferometer

(*a*) Schematic of FMCW Michelson sensor



(*b*) Frequency versus time for optical and electronic signals


Transmissive Mach–Zehnder fibre interferometers
FMCW Laser source
Detector

(*c*) Multiplexed FMCW arrangement (Al-Chalabi *et al* 1985)

Figure C3.1.28. The FMCW method of sensing and multiplexing, showing: (*a*) a schematic diagram of the FMCW Michelson sensor, (*b*) variation of frequency with time for optical and electronic signals and (*c*) an arrangement for multiplexed sensors using FMCW.

The technique of FMCW for sensor multiplexing suffers from several practical difficulties, when used with a large number of sensors. The main problems arise from the large number of optical paths, which can produce a background of unwanted signals, from both the expected sensor paths and additional stray paths due to multiple reflections [83]. Also, when it is desired to extract precise phase information from the method, problems may arise from nonlinearities in the laser frequency modulation ramp. Possible ‘pulling’ of the laser frequency by re-injection of retro-reflected light can be avoided by use of an isolator, but a semiconductor laser usually has intrinsic nonlinearities in its frequency chirp response due to the complex internal optical and thermal laser time constants.

Other problems arise because the technique necessarily requires an unbalanced interferometer for it to operate, so will eventually have reduced fringe contrast as the path differences increase, due to the finite coherence length of the source. It also necessarily suffers signal degradation from phase-noise, due to random frequency fluctuations of the laser source from the ideal saw-tooth, as, unlike certain other methods [67] the phase noise cannot be passively compensated for by using balanced optical paths. Limits on the permissible path imbalances greatly reduce the range of heterodyne frequencies that may be allocated for multiplexed sensors, a problem that may be exacerbated by nonlinearities in the frequency ramp and undesirable amplitude modulation of the source. However, in spite of some disadvantages, the FMCW technique is conceptually elegant and has the major advantage of being able to monitor steady-state changes of the length of an interferometric sensor, simply from the absolute magnitude of the heterodyne frequency.

More recently, a variation of FMCW has been used for addressing hydrophone arrays. Here it is used to generate a heterodyne carrier, and has been renamed the ‘phase-generated carrier’ method. Rather than modulating the source directly, a fibre-compatible PS, such as a mechanically strained (piezo-driven) coil, or an integrated optic FS can be used. These methods are discussed in detail in bibliography 1, vol 4, and in the excellent review by Kersey.

It should be pointed out that the FMCW method is not limited to situations in which the optical carrier is modulated. It has been shown by Mallalieu *et al* [84] that FMCW may also be operated using a frequency-modulated sub-carrier wave, rather than modulating the optical wave. This removes the need for both coherent sources and monomode fibre, and simplifies the optical system for sensors that do not require the use of optical interferometry, but because of the much lower frequency slew rates, will not normally achieve the phenomenal sensitivity to minute length changes that optical interferometry offers. It has excellent attractions for high-resolution interrogation of fibres, and improvements in the modulation of the source will be discussed again later in this chapter.

C3.1.4.8 Coherence multiplexing

The technique of coherence multiplexing is an optical method, which is similar in concept to a technique used in spread-spectrum RF communications. A signal, with a superimposed random (or pseudo-random) modulation, may be demodulated to recover information by correlating the received signal with a similarly encoded random (or pseudo-random), but delayed reference signal. In sensing systems, a broadband source of short coherence length may be used as a transmitted signal, as such a source will, in general, be subject to naturally occurring random phase or frequency excursions. If such a signal is guided via two equal, or near-equal, monomode fibre paths, then the signals suffer nearly equal delays, and have a strong correlation, provided the path difference is small compared with the coherence length of the source. Under these conditions, high-contrast interference fringes can be observed if the output signals are mixed on a square-law detector. If, however, the paths differ by very much more than the coherence length of the source, the fringe contrast becomes close to zero.

The arrangement in figure C3.1.29, first proposed by Brooks *et al* [85], shows how several remote Mach–Zehnder interferometers may be independently addressed using the coherence multiplexing technique, provided the optical path length differences, $l_1 - l_0$, $l_2 - l_0$ and $l_1 - l_2$, are all much greater than the coherence length of the source. The sensing method is based on observing the fringe shifts, which occur as the path differences $l_1 - l_0$ and $l_2 - l_0$, in the sensors 1 and 2, are changed by small increments, Δ_1 and Δ_2, respectively. The changes are observed by interferometric comparison with the corresponding fixed lengths $l_1 - l_0$ and $l_2 - l_0$ in the receiver interferometers. Only close-matched paths give visible interference fringes.

As a means of separately interrogating the outputs from a small number of remote sensors, the method has attractions. However, as the number of sensors increases, the number of possible optical paths in the network, from source to detector, increases very dramatically and the use of a very short coherence length source, such as an LED, soon becomes necessary. This then presents extreme difficulty in achieving the very close path-length equalization required, in order to ensure adequate fringe contrast, and results in a rather small dynamic range of measurement before fringe visibility is lost. Further practical disadvantages of the system are the need for each sensor to have a unique path length difference (and hence, without careful design, a different sensitivity) and the need for adequate

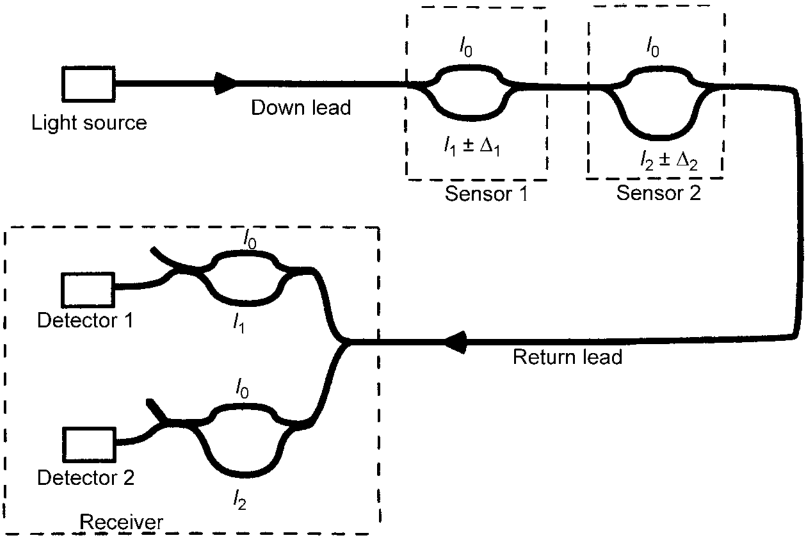


Figure C3.1.29. Schematic arrangement for coherence multiplexing, shown for the simplest case of separation of signals from only two remote Mach–Zehnder interferometric sensors.

stabilization of the receiver interferometers, to avoid undesirable errors due to drift or vibration in these reference interferometers.

In conclusion, therefore, it appears likely that, although elegant in conception, the method is less likely to find practical application in systems with more than perhaps 10 sensors, unless other multiplexing methods, such as WDM and TDM are also used.

A method for grating interrogation using coherence effects has been presented by Dakin *et al* [86]. Here, a number of in-line Bragg-grating-pair sensors are interrogated, in turn, using a scanned Michelson interferometer. Fringes are only observed when the path length of the interrogating Michelson interferometer matches the optical spacing of the Bragg gratings, and the rate of fringe crossing gives a measure of the wavelength of the gratings.

In the final section of this chapter, M Marcus will also give a detailed case study of work at Kodak on a white-light interferometer system for decoding multiple light paths in manufactured optical products. These have been used for assessing the inter-element spacing and aspects such as focal lengths of production cameras.

C3.1.4.9 Conventional sensors with an acoustic-sensing 'fibredyne' highway

As a final sensor multiplexing method, a hybrid technology will be described, which is really a non-invasive data collection highway. Its sensor use would be for collecting data from arrays of conventional electrical sensors, using a continuous optical fibre sensing highway. This highway is essentially a long, continuous, acoustic sensor, based on interferometric techniques, to which signals from electrical sensors could be coupled using piezoelectric transduction elements, held in close mechanical contact with the fibre or cable. In order to distinguish the separate signals applied to the common highway, it is possible to allocate electrical sub-carrier signals, of different frequency for each sensor, and allow the sensor output signals to modulate (intensity or angle modulation) these electrical carriers and impart information before they are applied to the piezoelectric transducers.

The original method, referred to as a 'fibredyne' system, was first devised and built by workers at UCL, London [87]. The first version (figure C3.1.30(*a*)) used a Mach–Zehnder twin-fibre interferometer arrangement, in which one fibre acted as a passive reference arm and the second was used as the strain-sensitive highway, which was subjected to an acoustic influence from the piezoelectric transducers. The method was initially proposed as a wideband fibre highway and had a high sensitivity, but, without use of a PC, would be subject to polarization fading when used with a normal monomode fibre.

A later version, from the same research team, used a single multimode fibre and relied on variation of the 'speckle' interference pattern emerging from the fibre, as the PZT transducer caused mode conversion in it (figure C3.1.30(*b*)). This method had the advantage of not requiring a separate reference fibre, but, as a result of the complex beating processes between many fibre modes, and the need for differential modulation of mode delays, it exhibited a somewhat lower sensitivity. It had a greatly reduced optical energy efficiency, as only a fraction of the speckle pattern was actually incident on the detector (otherwise only the *total* transmitted power would be detected, and phase modulation would not change this) and there was a strong possibility that either multimode fading or polarization fading could occur. The fading can only be prevented by arranging for diversity of phase, polarization and spatial arrangement of detecting the speckle. This requires multiple detectors and possibly active electronics to select non-fading channels. Despite these problems associated with this first

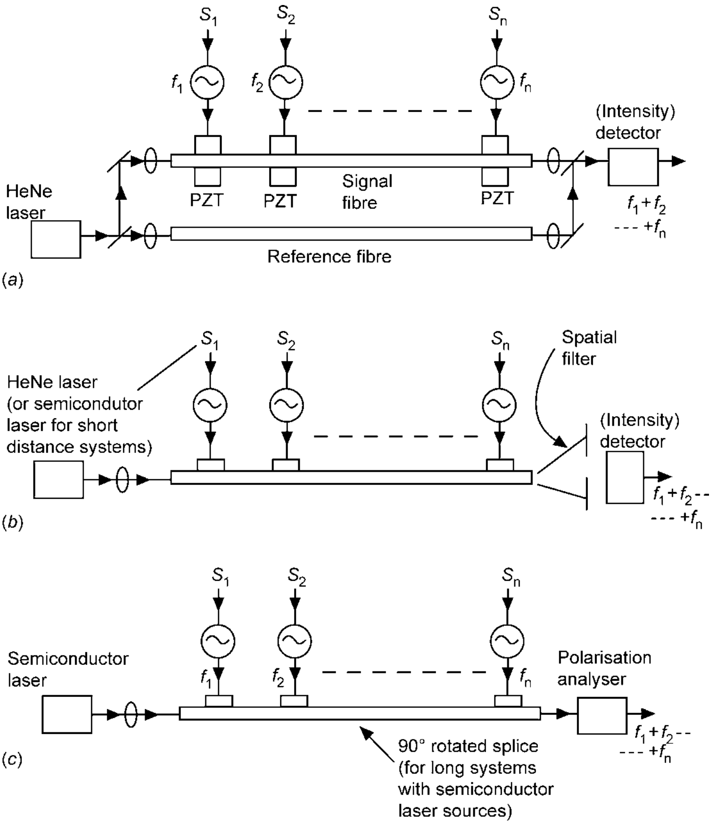

S_1
S_2
S_n
f_1
f_2
f_n
(Intensity) detector
HeNe laser
PZT
PZT
Signal fibre
PZT
f_1+f_2 --- $+f_n$
Reference fibre
(*a*)
S_1
S_2
S_n
Spatial filter
HeNe laser (or semicondutor laser for short distance systems)
(Intensity) detector
f_1+f_2-- --- $+f_n$
(*b*)
S_1
S_2
S_n
Semiconductor laser
f_1
f_2
f_n
Polarisation analyser
f_1+f_2-- --- $+f_n$
90° rotated splice (for long systems with semiconductor laser sources)
(*c*)

Figure C3.1.30. 'Fibredyne' highway systems: (*a*) original fibredyne concept using monomode fibre (Mach–Zehnder configuration), (*b*) multimode fibredyne highway (speckle modulation) and (*c*) polarimetric fibredyne highway (polarization modulation).

implementation [88, 89], a sensor telemetry highway was constructed and tested. In this, a number of stations could be fed in, using a set of channels, with relatively low radio frequencies for the carrier signals (≈1 MHz). The channels had an even lower baseband-channel frequency (information rate) of ≈10 kHz, which although low would be more than sufficient for many data collection requirements from sensors. This signal to highway coupling was carried out acoustically, as before, but, in these later systems, the acoustic waves were transmitted via the plastic protective sheathing of a fibre-optic highway cable, the reduction in coupling efficiency giving rise to a reduced information bandwidth.

An improved method has been devised (figure C3.1.30(*c*)), using a polarization-maintaining fibre, again with lateral acoustic excitation. If suitably aligned, this mechanical excitation has the effect of modulating the phase of one polarization mode, relative to that of the orthogonally polarized mode, resulting in the modulation of the polarization of the transmitted light [149]. This method is more efficient in its use of the received optical energy, as there are only two polarization modes in the fibre, and, when these interfere on the detector, no optical energy is wasted and fading can occur provided the correct phase bias is maintained between the two received polarization modes (this phase bias can be changed by stretchers to set the quiescent operating condition at the quadrature, or maximum-slope point of the sinusoidal response of the polarimetric interferometer).

C3.1.5 Distributed optical fibre sensors

We shall now discuss fully distributed sensing methods. Systems that permit the monitoring of not only the magnitude of a physical parameter or measurand, but also its variation along the length of a continuous uninterrupted optical fibre are particularly attractive, for both economic and logistical reasons. Distributed sensors not only allow a simple easily deployable sensing cable, where the communications link and sensor are one entity, but they also permit an easy and reliable comparison of a parameter at different points (the same interrogator takes measurements at different points) and the sensor cable measures at every point along the length with no 'dead spots'. This latter aspect is particularly important for safety related sensors (e.g. fire detectors or detectors for hazardous leaks of chemicals) and for sensors to detect intrusion through perimeters.

The simplest form of distributed sensor, the optical time domain reflectometer (OTDR) we discussed briefly earlier, has been in commercial use for many years as a measurement method for telecommunications, and is commercially available from many manufacturers. However, in spite of the relatively early development of the OTDR concept, it was several years before its use for the distributed measurement of parameters external to the optical fibre was envisaged [91]. Rogers suggested, for the first time, a method for the measurement of the spatial distribution of magnetic and electric fields, pressure and temperature, using OTDR concepts in conjunction with polarized light sources and polarization-sensitive receivers. More details of such sensors are given below.

C3.1.5.1 Backscattered sensors using the OTDR concept (general introduction)

All the sensors discussed in this section make use of radar-type (more accurately, LIDAR-type) backscattering (or backward-travelling light from other inelastic light generating processes, such as Raman, Brillouin or fluorescence) to make truly continuous measurements on unbroken optical fibres or fibre cables.

The basic method of optical time-domain reflectometry was, as described earlier (section C3.1.4.3, figure C3.1.24) the first type of distributed optical fibre sensor. In view of its importance here, we shall firstly briefly review it and then give more details of the principle of operation. A pulsed semiconductor laser is coupled into a section of fibre via a directional coupler, which serves also to couple the backscattered light fraction, captured and returned via the fibre on test, to the high-speed optical receiver.

In normal telecommunications fibres, the Rayleigh component of the scattered light represents well over 98% of the returning signal (except during the short time intervals when more intense specularly reflected pulses return from discrete discontinuities, such as connectors, air-spaced splices or distal fibre ends).

For uniform fibres, the detected temporally varying Rayleigh scattered power $I(t)$ varies as the product of at least five important physical factors, all of which can either be chosen by appropriate design or component selection. These are firstly the launched energy, E_0, secondly the scattering attenuation coefficient, α_s, in the fibre, thirdly the fraction, S, of scattered light that is captured by the fibre in the return direction, fourthly the inverse-4th power λ^{-4} of the optical wavelength, λ, and finally the two-way optical attenuation factor, $\exp\left(\int \alpha(x)\,\mathrm{d}x\right)$, that occurs during propagation from the source to the scattering point and back to the detector:

$$I(t) = \frac{1}{2} E_0 v_g \alpha_s S \exp\left(\int \alpha(x)\,\mathrm{d}x\right).$$

There are also a few other factors that are more difficult to change significantly, such as v_g, the velocity of light in the guide, which has to be close to that of silica, as the doping levels are low, and the fixed numerical factor of 0.5. The factor S has a close to quadratic dependence on the numerical aperture of the fibre, and hence on the core/cladding index difference, so is higher for high numerical aperture (NA) multimode fibres. However, the significantly lower attenuation coefficient, $\alpha(x)$, in monomode fibres can become a more important advantage when very long lengths of fibre are probed, despite their much lower S value. The SNR of an OTDR reduces very rapidly as the distance resolution improves, as firstly much shorter pulses are needed (reducing the launched energy, E_0) and secondly a high speed receiver has a wider noise bandwidth and usually a higher noise spectral density too, both conspiring to give far worse noise performance with short pulse excitation and fast detection[†]. Because the scattering coefficient in high-quality fibres does not usually vary significantly along the length, the method has proved extremely useful for measuring spatial variations of fibre attenuation. However, if the geometry or numerical aperture of the fibre varies significantly, changes in the guidance properties of the fibre (primarily in the modal 'V' number, that changes the number of allowable modes) will cause additional variations in the backscattered signature [92, 93]. These workers showed, however, that variations may be at least partially compensated for, provided two separate OTDR signatures are taken from opposite ends of the same fibre. Clearly, this requires two instruments or a loop sensor with a two-way fibre switch.

In the following subsections, several sensing methods using OTDR, and variations based on various inelastic scattering and fluorescent methods, will be described.

C3.1.5.2 Monitoring of variations in attenuation using OTDR

The original objective of the OTDR method was to examine attenuation variations in manufactured and installed lengths of optical fibre. One of the first suggestions that it may be possible to construct distributed sensors, using the attenuation characteristics observed was made by Theocharous [94], who proposed a method for the measurement of temperature. If an OTDR return signal is differentiated with respect to time and normalized by division by the instantaneous value of the signal, a measure of the fibre attenuation is obtained. If, therefore, a fibre with temperature-dependent attenuation is used, variations of temperature along the length may be monitored.

[†] Measures to reduce this will be discussed later, but one useful way to improve an OTDR system is to include optical fibre amplifiers to boost the launch power and to act as an optical pre-amplifier for returning signals.

The technique appears at first sight to be highly attractive, but a simple semi-qualitative analysis suggests that the number of distance-resolution elements is likely to be low. To measure the average temperature within each distance-resolution element, an accurate loss measurement is necessary, in order to determine the smaller temperature-dependent changes in loss. Because the OTDR return is usually rather noisy, a temperature-dependent loss of at least 1 dB in the resolution element (whatever its length) will probably be necessary, in order to accurately detect small changes, say of 1% (or 0.01 dB) in this loss. Thus, the round-trip attenuation in this example will be a loss of 2 dB for each resolution element, times the number of such elements, giving a likely limit to the number of measured resolution elements of perhaps ten, particularly bearing in mind the poorer signal/noise ratio that will return from the later sections of fibre.

A comprehensive study of rare-earth-doped fibres, showing how they could be used to determine the line-averaged temperature over their length, was carried out by Quoi *et al* [95]. Many different rare-earth fibres were considered, and the interesting concept of detecting light at two wavelengths, one where attenuation increased with temperature, and another where it decreased, was introduced as a means of compensating (by ratioing the two detected signals) for losses in bends, splices or connectors.

Because of the attenuation problem over longer lengths that was discussed above, distributed sensing methods monitoring attenuation are most likely to have practical application for sensing using fibres or cables which have a low intrinsic loss, but which then suffer significantly increased attenuation at just one point or small region due to some event it is desired to detect or locate. Examples of this are largely safety or security related, such as the detection (and location) of fire using a cable that might have high loss if it becomes hot. Several published applications for other applications will now be discussed below.

An early means of using OTDR monitored attenuation variations for distributed radiation dosimetry was presented by Gaebler and Braunig [11]. In this application, a short section of a fibre, which was exposed to ionizing radiation, suffered excess attenuation, enabling simultaneous detection and location of the radiation exposure (see figure C3.1.31). If used as a sensor for the monitoring of

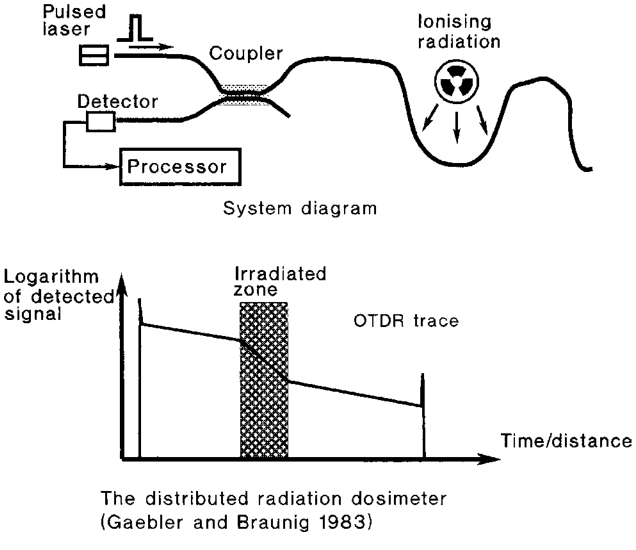


Figure C3.1.31. Schematic of a distributed radiation dosimeter, using radiation-sensitive fibre, with an OTDR to detect regions of increased loss. There is potential for leak-detection system for radioactive materials.

exposure over a short section of a more extensive overall length, the quiescent attenuation before radiation is merely that of a normal low-loss communication fibre, enabling the increased attenuation of the short exposed section to be measured with reasonable accuracy. The method would be attractive for detecting leaks of highly radioactive material, but silica fibres are rather insensitive, and a large part of the radiation-induced loss recovers quickly with time. Lead-glass fibres are more sensitive, but have higher intrinsic losses.

Several other loss-modulation effects in fibres, which were discussed in the first part of this chapter, are also amenable to OTDR interrogation. The first of these is the plastic-clad silica fibre system (see earlier figure C3.1.6) of Pinchbeck and Kitchen [12] where the fibre showed increased attenuation when cooled, due to increase in the refractive index of the cladding polymer. Clearly, the OTDR offers the possibility of leak location, whereas that shown in figure C3.1.6 can only offer detection of leaks, as it only permits simple detection of fibre loss (figure C3.1.32).

The second we shall mention is the commercially available ‘Herga’ fibre cable (figure C3.1.3), which exhibits high microbending loss when subjected to lateral pressure, as a result of its novel spiral plastic sheathing arrangement. Figure C3.1.33 shows a schematic of an OTDR-based distributed sensor, using a pressure sensing cable of this type.

Sensors such as these have been proposed by Alan Harmer, of Battelle research, for use as a distributed sensor to detect intruders, by monitoring losses from pressure of the intruder’s foot on the cable.

The same system layout of figure C3.1.33 can clearly be used with the water ingress sensor described earlier ([9, 10], see also ‘Microbend sensors’ section and figure C3.1.4), which is based on the same microbending concept, and indeed this sensor type was developed by the authors using an OTDR instrument to detect the changes. One useful application is to detect leakage of water in telecommunications ducts, as this can cause damage to optical cables, particularly if it generates hydrogen, or if it freezes and breaks the fibres.

Another useful intrinsic loss mechanism that can be used is that of evanescent field absorption in the fibre cladding (figure C3.1.34). Although, in principle, a bare unclad fibre could be used to detect absorbing material directly, this would easily be contaminated. It is sensible therefore to use a polymer cladding with suitable properties. Many polymer cladding materials will adsorb oil or other liquids

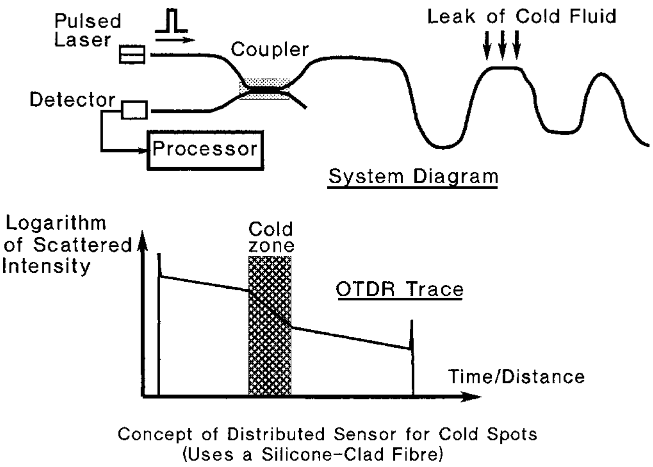

Pulsed Laser
Leak of Cold Fluid
Coupler
Detector
Processor
System Diagram
Logarithm of Scattered Intensity
Cold zone
OTDR Trace
Time/Distance
Concept of Distributed Sensor for Cold Spots
(Uses a Silicone-Clad Fibre)

Figure C3.1.32. Schematic of a distributed leak-detection system for cryogenic liquids, using PCS fibre, with an OTDR to detect regions of increased loss (see also figure C3.1.6).

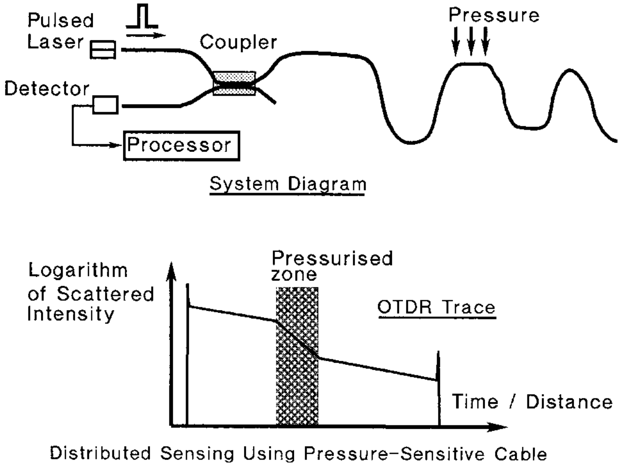


Figure C3.1.33. Schematic of a distributed detection system for lateral pressure, arising, for example, from an intruder treading on cable. Sensor uses pressure-sensitive cable of the type shown in figure C3.1.3, with an OTDR to detect regions of increased loss.

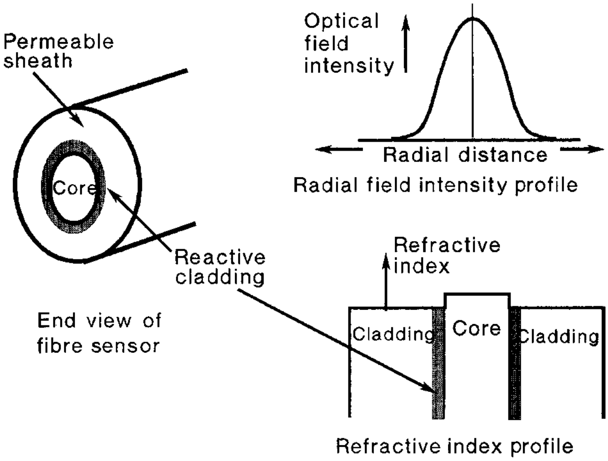
and change their absorption [96] but if chemical indicators are available, and can be incorporated in the polymer, they can offer a more selective response to a desired target species. Blyler *et al* [97] have made a sensor for measurement of ammonia using reactive cladding material.

It should be noted that it is not just absorption that may be used to form the basis of the sensor. The use of fluorescent coatings will be discussed later, after discussion of the use of other types of scattering processes in sensors.

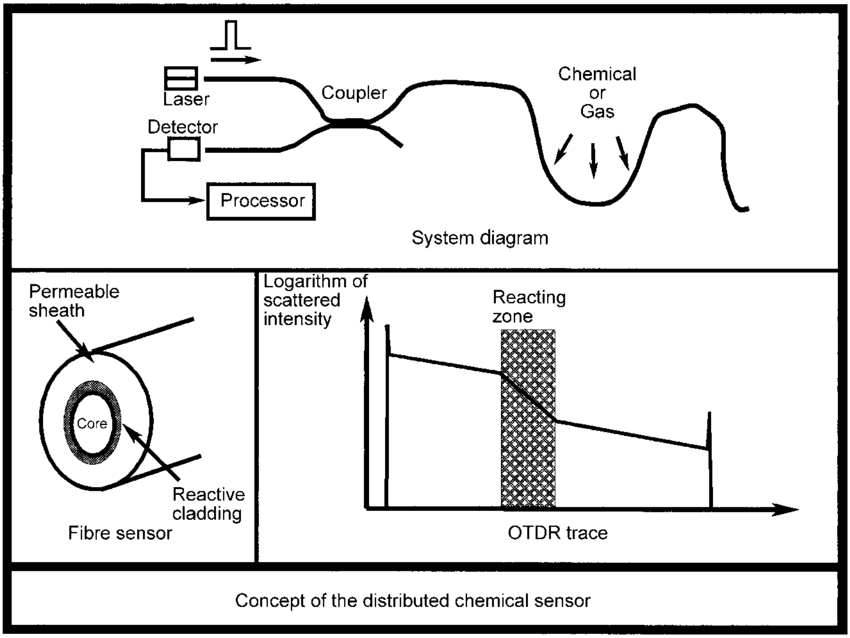
C3.1.5.3 Variations in Rayleigh backscatter characteristics

As already discussed, the use of OTDR to monitor fibre attenuation depends on the constancy of the Rayleigh backscattering coefficient along the length of the fibre. However, this may vary significantly in two basic ways, even in fibres of uniform geometry and composition. The first form of variability occurs in monomode fibres, using polarized illumination and polarization-sensitive detection. An arrangement as in figure C3.1.24 is used, with a polarized light source, and now a polarization analyser is used at the detection end. This diagnostic method, known as polarization optical time-domain reflectometry (POTDR), relies on the high degree of preservation of polarization exhibited by Rayleigh and Rayleigh–Gans scattered light in silica fibres, leaving the polarization changes that occur due to two-way propagation in the fibre itself to be observable.

The POTDR method was first suggested by Rogers [91], who pointed out its potential for distributed measurements of magnetic field (via Faraday rotation), electric field (via the Kerr quadratic electro-optic effect), lateral pressure (via the elasto-optic effect) and temperature (via the temperature dependence of the elasto-optic effect). The first experimental measurements were reported by Hartog *et al* [98], who used the technique for a distributed measurement of the intrinsic birefringence of a monomode fibre, and Kim and Choi [99] who measured the birefringence induced by the bending of a wound fibre. Ross [100] carried out the first measurement of a variable external field, using the Faraday rotation of polarization as an indication of the magnetic field environment of the fibre. A comprehensive theoretical treatment of the POTDR method has been presented by Rogers [101].


Permeable sheath
Core
Reactive cladding
End view of fibre sensor
Optical field intensity
Radial distance
Radial field intensity profile
Refractive index
Cladding
Core
Cladding
Refractive index profile

(*a*) Schematic of structure of chemical–sensing fibre.
(Also shows variation of refractive index and optical field intensity)


Laser
Detector
Coupler
Chemical or Gas
Processor
System diagram
Permeable sheath
Core
Reactive cladding
Fibre sensor
Logarithm of scattered intensity
Reacting zone
OTDR trace
Concept of the distributed chemical sensor

(*b*)

Figure C3.1.34. Schematic of a distributed chemical sensor system, using fibre with a reactive cladding that changes loss after exposure to the chemical. The structure of the fibre, the mode field pattern and the refractive index profile are shown in (*a*). Possibilities include permeation of a polymer cladding, by, for example, oil leaks, affecting either the refractive index or absorption properties, or use of a cladding with a chemical indicator designed to be selective to a desired target chemical or chemical species. The OTDR response when detecting leaks is shown in (*b*).

On first consideration, the POTDR technique appears to be attractive for the measurement of a large number of parameters. However, its main drawback, as with many other potentially useful sensing methods, is the variety of parameters to which it can respond, the sensitivity to strain and vibration being particularly troublesome when it is desired to measure other things. In addition, POTDR requires the use of monomode fibres, which can, when used with narrow linewidth laser sources, have particular problems due to coherent addition from multiple Rayleigh backscattering centres [102].

Moving on now from considering variations in the polarization of backscattered light, the scalar magnitude of the Rayleigh scattering coefficient may vary with temperature in some types of fibre. If so, a simple OTDR arrangement, now preferably without polarization sensitivity, can be used to perform distributed temperature measurements. Unfortunately, in normal vitreous fibres, of silica or multi-component oxide glasses, the majority of backscattered energy arises from Rayleigh or Rayleigh–Gans scattering from frozen-in refractive index variations. These are tiny regions where high-temperature thermally induced density changes were ‘frozen’ into the glass structure as it is cooled from the melt. As a result, once the glass has cooled, the scattered light intensity varies very little with the temperature of the glass. In many normal room-temperature liquids, however, the scattering arises from real-time thermodynamic fluctuations in the refractive index which are dependent on the ambient temperature and therefore the scattering will now show a significant temperature coefficient.

This effect has been exploited for distributed sensing of temperature, in the system described by Hartog and Payne [103]. Unfortunately, the use of liquid-filled fibres presents difficulties, which, at best, complicate the method and, under some circumstances, restrict its use:

(i) The use of liquid-filled fibres is inconvenient because of the need for expansion reservoirs.

(ii) There are temperature range restrictions imposed by the freezing and boiling points of the liquid.

(iii) If the central region of a long length of such a fibre is rapidly cooled, it is possible to create voids by rapid thermal contraction, before pressure differentials can refill the tube against the resistance of viscous forces.

(iv) Impurities or dust particles may increase the scattering cross-section or increase the localized loss of the fibre, giving the appearance of false temperature variations in the OTDR return.

(v) The numerical aperture of the liquid-filled fibre shows significant temperature dependence, generally reducing as the temperature is raised, allowing hot zones in particular, to affect the calibration of results further along the fibre. (In the absence of mode-conversion, this effect may be reduced by the use of mode filters.)

However, in spite of the above potential drawbacks, which are not uncommon with initial ground-breaking developments, the system worked well over moderate temperature ranges and was capable of monitoring the temperature distribution to ≈0.2°C resolution, over several hundred metres of fibre, with a distance resolution of the order of 2 m. This was a remarkable advance, considering that this was the first experimentally demonstrated temperature profile measuring method for use with optical fibres.

C3.1.5.4 Distributed anti-Stokes Raman thermometry (DART)

If the spectral variation of backscattering from a germania-doped silica fibre is examined (figure C3.1.35), it may be seen that there is a strong central line, primarily due to Rayleigh (or Rayleigh–Gans) scattering, but which also contains a weaker (spectrally unresolved in figure C3.1.12) contribution from Brillouin scattering. At each side of the central line, however, there are side-lobes due to Raman scattering. These may be used to detect temperature profiles in conventional vitreous communications fibres using modified Raman OTDR techniques [104, 105].

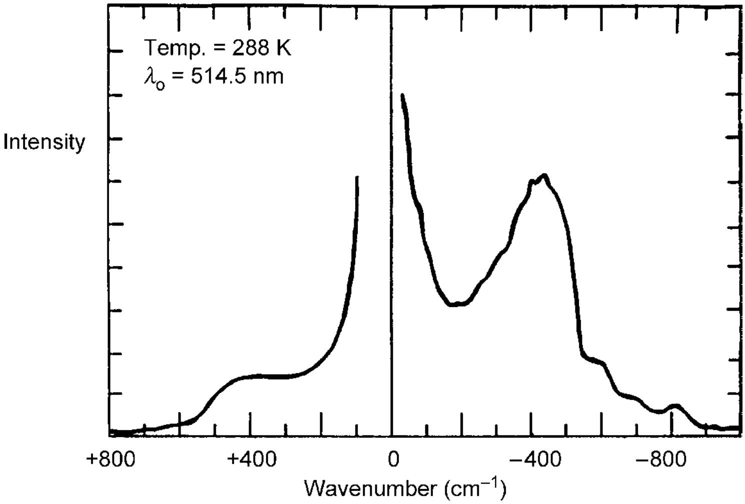


Figure C3.1.35. Typical spectrum of Raman backscattered light in germania-doped silica fibre (measured by N Ross, then at CERL).

A schematic of a basic Raman temperature sensor system is shown in figure C3.1.36. The system is essentially a dual-wavelength OTDR system, where a WDM filter is used to select out the two Raman bands and direct the signals to two detectors. The elastic Rayleigh scattering of the strong incident light signal (central peak in figure C3.1.35) has to be removed in the filtering process. A signal processor averages the time-varying returns, and then takes the two separate OTDR results from each detector channel, divides the two signals coming from each selected point along the fibre to determine the Raman ratio, and hence determine the temperature at each point.

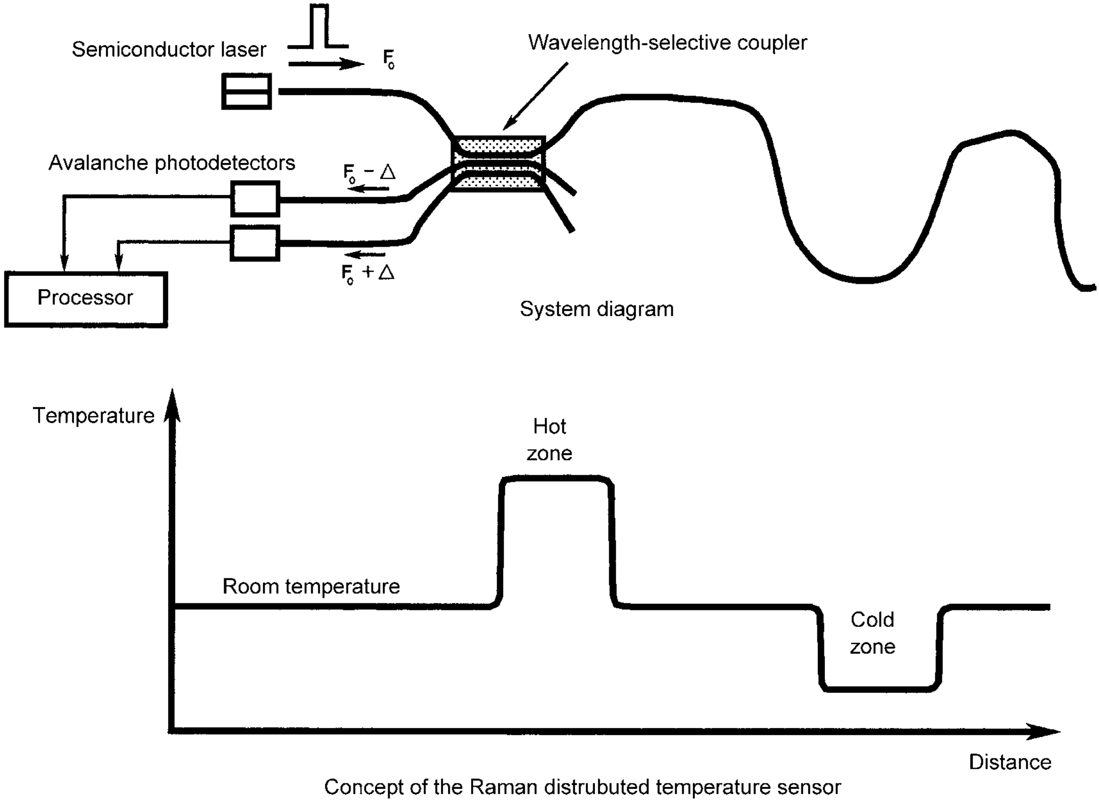
From standard texts on Raman scattering, the temperature-dependent ratio, $R(\theta)$, of the anti-Stokes (higher-frequency band) and Stokes (lower frequency band) scattered intensity, at wavelengths, λ_a and λ_s, respectively, and assuming equal frequency separation from the central excitation laser line, is given by the relationship:

$$R(\theta) = (\lambda_a/\lambda_s)^4 \exp(-hc\nu/KT)$$

where h is Planck's constant, c is the velocity of light *in vacuo*, K is Boltzmann's constant, θ is the absolute temperature and ν is the frequency of the incident light.

Therefore, in addition to the distance information provided by the time delay of returning signals, a measurement of the ratio of Stokes and anti-Stokes backscattered light in a fibre can, in principle, provide an absolute indication of the temperature of the medium, irrespective of the light intensity, the launch conditions, the fibre geometry and even the composition of the fibre. In practice, however, a small correction will usually need to be made for the difference in the fibre attenuation between the Stokes and anti-Stokes wavelengths, and if convenient, it may be desired to use the dual-end measurement method discussed earlier to compensate for fibre non-uniformities.

The Raman technique appears to have only one significant practical drawback: that of a very weak return signal, the anti-Stokes Raman-scattered signal being between 20 and 30 dB weaker than the Rayleigh signal, which itself is already typically 50 dB weaker than the incident light. In order to avoid an excessive signal averaging time, measurements have been taken using relatively high launched powers from pulsed lasers, and extensive signal processing is performed to average the signals. In the first experimental demonstration of the method [104], a pulsed argon–ion laser was used in conjunction with an early telecommunications-grade 50/125 μm GRIN fibre. Dakin *et al* [105] described subsequent



System diagram

Concept of the Raman distrubuted temperature sensor

Figure C3.1.36. Basic arrangement of distributed temperature sensor using Raman backscatter. Range is determined by two-way time of flight of light and temperature by the calculation of the ratio of Raman Stokes and anti-Stokes signals at each range.

results from a considerably more compact system, based on a far more practical and convenient semiconductor laser source. A result from an early commercial prototype constructed by the group of A Hartog (previously at York Ltd), is shown in figure C3.1.37. Here the result has been normalized to remove the effects of fibre attenuation, and clearly shows the hot and cold regions of fibre.

Since those early results, the Raman OTDR system has reached commercial maturity, and is another success story in the field of fibre sensors. The system has been manufactured for several years by York sensors, by the Hartog group, which is now part of the large Schlumberger company, and in Japan Hitachi cable have demonstrated engineered systems. Most systems have the capability to address multiple channels via optical fibre switches, and can operate over many km of fibre. Significant signal gains have been made using Q-switched fibre lasers to launch more light, but there are launch-power limits in monomode fibres due to the onset of stimulated Raman processes. Major applications have included fire detection, oil-well logging, chemical process plant and furnace measurements and in-line monitoring of high voltage cables and other electrical plant, although there clearly many more potential applications may become economically viable as costs become lower.

C3.1.5.5 Time-domain fluorescence monitoring

The re-emission spectrum of most fluorescent materials generally exhibits a significant temperature variation. Thus, if an optical arrangement similar to that used in figure C3.1.36 for Raman OTDR is

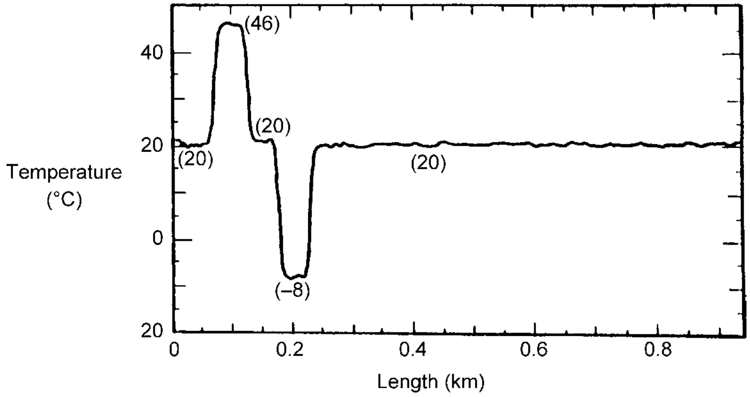


Figure C3.1.37. Early Raman temperature profile result, from York Ltd.

constructed, with a laser exciting source as before, and with the detector filters now selected to examine regions of the fluorescent decay spectrum having the maximum possible differential temperature variation, a distributed temperature sensor should be possible. The potential attraction of the method, first proposed by Dakin [106], is that the fluorescent quantum efficiency may be many orders of magnitude higher than that for Raman scattering and higher doping levels may greatly enhance the signals in short distributed sensor systems. However, there remains a problem with the availability of suitable fibres.

Silica-based optical fibres, with rare-earth dopants giving high fluorescent efficiency, have been prepared [107], but sensors using these give very poor distance resolution due to the long fluorescent lifetimes. It is possible to reduce fluorescent lifetime by using materials with increased coupling to non-radiative processes, but this unfortunately also reduces the fluorescent efficiency. Polymer fibres may perhaps offer more promise in short distance systems, as these may be doped with organic dye materials with an excellent combination of high quantum efficiency (~50% or better) and fluorescent lifetimes of the order of only a few nanoseconds.

Dakin and Pratt [90] made a theoretical comparison between distributed temperature sensors based on the techniques of temperature-dependent absorption, scattering, Raman scattering and fluorescence. It was predicted that, although doping with strongly fluorescent materials will necessarily increase the loss in an optical fibre, this should, for short distance operation, be more than offset by the much higher fluorescent light levels theoretically attainable. To make such sensors in practice, it will be necessary to obtain fibres having short-lifetime fluorescent dopants showing the desired thermal variation.

Lieberman *et al* [108] has reported a distributed chemical sensor for oxygen, using a fluorescent polymer cladding coating on a fibre that has its fluorescence efficiency quenched by oxygen. A low-index siloxane cladding, doped with a 9,10-di-phenylanthracene dye, was used. The evanescent field of the incident radiation coupled into the cladding dye and excited the fluorescence, and the fluorescence was coupled back, somewhat more weakly of course, by an inverse coupling process, to excite forward and backward guided modes of the fibre. Although a distributed sensor, in that continuous measurements over an extended length were taken, the method does not yet appear to have been used to determine the *variation* in oxygen concentration over the fibre length. Clearly, to do this, a dye with a short fluorescent lifetime would be necessary, and, as with Raman OTDR, significant signal averaging would probably be necessary to detect the weakly coupled signals.

C3.1.5.6 Distributed sensors using spontaneous backscattered Brillouin light

The Raman DART system described above has evolved to reach commercial maturity relatively quickly, primarily because of its basic optical simplicity, its simple telecommunications components requirements, and because of its similarity to the already-mature conventional OTDR method. However, Raman signals are extremely weak compared to Rayleigh elastically scattered light, so relatively long electronic integration times (typically tens of seconds) are needed to achieve the necessary signal/noise ratio to realize temperature resolutions of 1°C or better, particularly if it is intended to probe lengths of fibre above 5 km or so.

The total light level associated with spontaneous Brillouin scattering is typically two orders of magnitude stronger than that of Raman light, the only disadvantage being the very close frequency spacing of this light to the incident excitation line. Because Brillouin light occurs due to scattering from relatively low-energy acoustic phonons, the photon energy change is very small. An alternative simple classical physics viewpoint, which also predicts the correct optical frequency shift, is to consider that the scattering arises from thermodynamically induced moving acoustic waves in the core of the glass, so they become Doppler shifted. The resulting frequency shift in the light is therefore much smaller than that for Raman scattering, being typically only of the order of 12 GHz, so very narrow band interferometric filters are necessary to perform optical separation. Suitable narrowband filters are interferometric filters. These may be of the Fabry–Perot type, all-fibre types, based on the Mach–Zehnder or Michelson configurations, or equivalent interferometers in integrated optics form.

An alternative is to use coherent detection, heterodyning the Brillouin light with an optical local oscillator, suitably shifted from the incident laser to give a conveniently low beat frequency to process electronically. A local oscillator with a very stable frequency offset can be derived by frequency shifting a portion of light from the main excitation laser (pump laser). The signal at the detector is then a low frequency beat signal that exhibits a large fractional change for small changes in the Brillouin backscatter frequency. Once this frequency separation has been performed, a number of very useful sensor types can be constructed. The method of Kurashima *et al* [151, 155] allowed accurate determinations of distribution of mechanical strain in fibres from the Brillouin frequency shift. The heterodyning process makes it relatively easy to measure the Brillouin frequency shift, ν_B, which is a function of the fibre core refractive index, n, the acoustic velocity, v_s, in the fibre core and the incident optical wavelength, λ:

$$\nu_B = n v_s \lambda^{-1}.$$

ν_B is typically of the order of 12 GHz, and, according to various papers by Horiguchi *et al* [109–112], varies with both temperature (typically of the order of 9.4×10^{-5} fractional change per degree kelvin) and strain ($\sim 4.6\epsilon$ change, where ϵ is the tensile longitudinal strain).

In addition to the frequency shift, a convenient temperature-dependent intensity relationship that can be used [113] is the ratio of Rayleigh scattered intensity to total Brillouin intensity, or the Landau–Placzek ratio R_{L-P}. For a single component glass, the ratio, R_{L-P}, of Rayleigh scattered light to total Brillouin scattered light is given by:

$$R_{L-P} = T_f / T(\rho_0 v^2 \beta_T - 1)$$

where T_f is the fictive temperature of the glass, T is the absolute temperature, ρ_0 is the density, v is the acoustic velocity and β_T is the isothermal compressibility of the melt at the fictive temperature. Clearly, unlike the Raman ratio discussed before, this new ratio will depend strongly on the fibre core material properties, so renewal of a fibre by one of different type or composition is more likely to create the need for system re-calibration. However, because the signal/noise ratio can be much higher in Brillouin

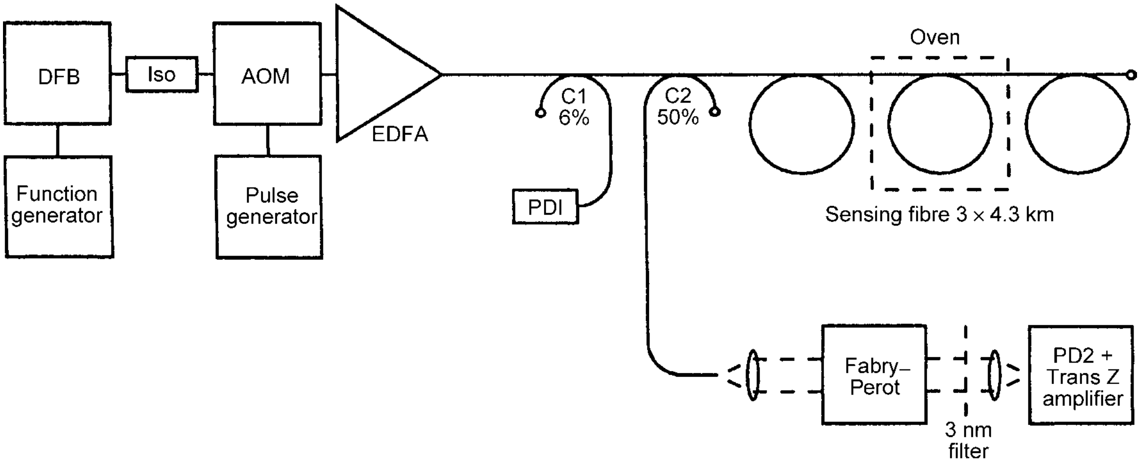


Figure C3.1.38. The distributed temperature sensing system of Wait and Newson, using measurement of the Landau–Placzek scattering ratio (the ratio of total Rayleigh scattered light to total Brillouin light).

systems, there is now scope for measurement over a much longer range. A simple arrangement for measuring temperature using the Landau–Placzek ratio method [113] is shown in figure C3.1.38.

Microwave-frequency heterodyning can be used to detect the frequency shift of the Brillouin light. The frequency shift, which is generally a stronger function of strain than temperature, can be monitored, and the intensity signals can be used as a more temperature-dependent quantity, giving the possibility of distributed sensing of both strain and temperature. Maughan *et al* [114] carried out simultaneous measurements of both these quantities over 30 km of fibre. It was possible to measure distributed temperature profiles over a length of 57 km. A strain resolution of 100 $\mu\epsilon$ was achieved, and a temperature resolution of 4 K.

More methods using spontaneous Brillouin scattering with optical frequency domain processing will be discussed below. The alternative nonlinear process of stimulated Brillouin scattering will be discussed later, when systems involving interaction of counter-propagating light beams are described.

C3.1.5.7 The optical frequency domain reflectometry (OFDR) technique

This distributed sensing method is essentially similar to the FMCW technique we discussed earlier as a multiplexing method. An FMCW system can be operated in OTDR-like backscattering mode in a continuous length of monomode fibre [115, 116]. Now, a portion of the launched light signal, derived from the optical source with a fibre splitter, is added as a local oscillator, to coherently detect returning signals, and a beat frequency is obtained which increases in frequency as a function of the distance to the point in the fibre from which the light was backscattered. If the detected beat signal is displayed on a conventional electronic spectrum analyser, the detected power within each small frequency interval represents the scattered light received from the section of fibre that is situated at a distance corresponding to the frequency offset observed. As the frequency slew rate of current-ramp-driven semiconductor laser diodes may be very high (100 GHz s^{-1} is easily achievable) and the frequency resolution of commercial electronic spectrum analysers is a few Hz or less, the technique can have a far superior distance resolution capability than conventional OTDR methods. Kingsley and Davies [116] even suggested use of the technique for distributed measurements over the very small scales involved in integrated optical waveguide circuitry, where resolutions as low as a cm, or often much less, are required.

A major potential problem with OFDR is the coherence function of the source, which will modulate the received spectrum and therefore distort any spatial variation of scattering that it is desired to observe. Another approach to the problem uses, as with the FMCW method discussed earlier, a frequency-modulated sub-carrier to amplitude modulate the source [117]. This removes the problems due to source coherence, but presents a reduced resolution due to the lower frequency slew rate possible with electronic sub-carrier systems.

The FMCW technique still shows limitations when it is desired to provide high sampling rate information, or to achieve a high spatial resolution of about cm order. The method we will describe next is an improvement on this basic method.

C3.1.5.8 Application of the 'synthesis of optical coherence function' method for high resolution distributed sensing, using elastic (Rayleigh) scattering

An alternative technique to synthesize interference characteristics arbitrarily, called the 'synthesis of optical coherence function', or SOCF for short, has been proposed and extensively developed by K Hotate of RCAST, Tokyo [118–120]. In this technique, the frequency of a laser light source, connected to an unbalanced fibre interferometer, is modulated, using an appropriate electrical bias waveform, and the phase of a lightwave propagating in just one arm of the interferometer is also modulated synchronously with a similar waveform. With this method, any arbitrary shape of coherence function can be synthesized, hence setting the amplitude of the interference fringe in the interferometer to be a function of the path length difference between a lightwave returning from the remote sensing point and that returning from another defined (reference) point.

In recent years, to assist with installation and maintenance of optical fibre subscriber networks, very precise measurements of the spatial distribution of reflective discontinuities (e.g. fibre breaks or connectors) are required. For applications such as monitoring repeater stations, it is desired to measure at a distance of several km, with typically a few cm or less spatial resolution, a specification far too difficult to achieve with conventional time-domain methods. The standard FMCW sensing method described above, with its high spatial resolution and wide dynamic range, could be a possible candidate, but its measurement time must be made much shorter than at present possible, as otherwise the optical phase noise arising from environmental fluctuation is a problem, particularly over a long fibre.

As a more viable alternative, a solution has been developed using the synthesis of optical coherence function method [118, 119, 121]. Figure C3.1.39 shows the experimental arrangement. A switch generates a series of wide optical pulses. The optical pulse forms a window to select and determine the desired 'range gate', or the boundaries of the chosen test region along the fibre under interrogation. This is because a pulse from the reference path can only overlap with a pulse from the path of the tested fibre when the two paths have nearly the same lengths. A coherence function having delta-function-like peaks can be synthesized using the FM waveform shown in figure C3.1.39.

The coherence function is synthesized, in such a way that, firstly, only one coherence peak can correspond to the region under test, and, secondly, so that all other undesired periodical coherence peaks are masked by the time window. The desired peak position is then achieved by scanning, to modify the phase modulation waveform. Consequently, the distribution of reflectivity within the narrow selected time/range window can be measured with excellent temporal, and hence spatial resolution. In order to measure in a new range gate or window, the reference delay is changed to select, and determine, this new region. An example of the reflectivity distribution obtained by this system is shown in figure C3.1.40.

Two reflections from optical connectors, of magnitude around −30 dB, can be seen clearly, while a strong reflection from the far end of the fibre is completely suppressed by the pulse window. Even when

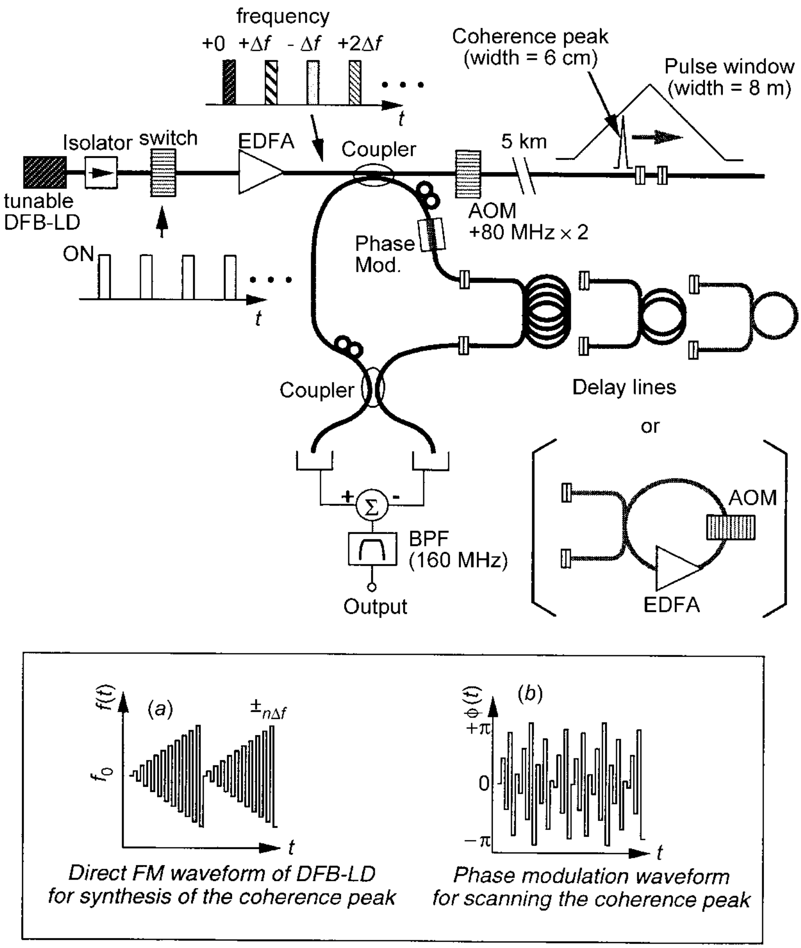

frequency
+0 +Δf - Δf +2Δf
t
Coherence peak
(width = 6 cm)
Pulse window
(width = 8 m)
Isolator
switch
EDFA
Coupler
5 km
tunable
DFB-LD
AOM
+80 MHz × 2
Phase
Mod.
ON
t
Coupler
Delay lines
or
AOM
EDFA
+
Σ
-
BPF
(160 MHz)
Output
f(t)
(a)
±nΔf
f0
t
Direct FM waveform of DFB-LD
for synthesis of the coherence peak
ϕ(t)
(b)
+π
0
−π
t
Phase modulation waveform
for scanning the coherence peak

Figure C3.1.39. Schematic of distributed sensing using the 'synthesis of optical coherence function' method conceived and developed by K. Hotate.

measuring through 5 km of optical fibre, the spatial resolution can be as good as 6 cm, using a range-gate region of 8 m. Clearly, this range gate can be moved to select and measure in any chosen region of fibre.

C3.1.5.9 The transmissive FMCW method for disturbance location

The FMCW methods may be used to locate discrete points where mode coupling in a fibre has occurred, provided the fibre is capable of supporting two modes (e.g. different polarization modes in a high birefringence fibre) having significantly different phase velocities. External disturbances, which cause cross-coupling from the initially excited single mode, will mix or beat on a suitable detection arrangement situated at the far end of the fibre. The beat signal will have a frequency dependent on the distance from the source, at which the coupling to the second mode has taken place.

This approach, first suggested by Franks *et al* [122], is depicted in figure C3.1.41. This particular implementation used a birefringent fibre, with all the transmitted signal energy being launched into only one of the two principal polarization modes of the fibre. The disturbance to be monitored was a lateral pressure on the fibre cable, which caused coupling of light into the orthogonal polarization mode.

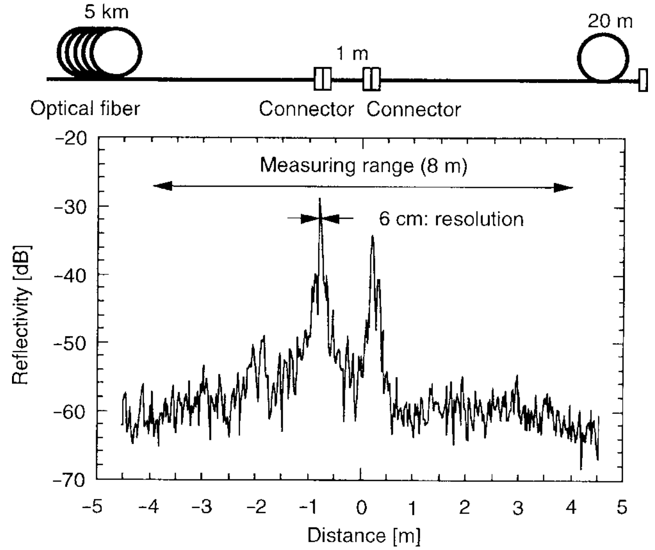


Figure C3.1.40. Trace showing backscatter signals from a pair of 5-km distant connectors, in a simple fibre network shown at top of figure. This was taken using the 'synthesis of optical coherence function' method of K Hotate.

A convenient attribute of the technique is that the relatively close velocity matching between the polarization modes, even when using the so-called high-birefringence fibre, allows FMCW techniques to be operated over lengths very much longer than the instantaneous coherence length of the source. Two potential difficulties exist with the method, however. Firstly, mechanical strains of certain critical magnitudes may cause coupling of power from one polarization mode to the other and then completely back again, resulting in no net beat signal. Secondly, disturbances that occur exactly in the direction along a fibre polarization axis will cause no mode coupling. Otherwise, except for these somewhat unlikely conditions, the technique appears a simple and elegant method of locating the position of lateral disturbances on a continuous fibre, but is probably unsuited to measuring the magnitude of the disturbance.

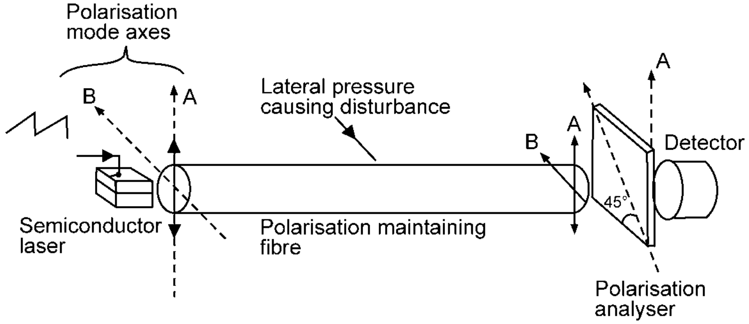


Figure C3.1.41. Transmissive FMCW disturbance location sensor [122]. This takes advantage of the different propagation speeds of the two principal polarization modes in a birefringent fibre, in order to provide the time delays needed for generation of the FMCW beat signals at the detector. There is no beat signal until pressure causes coupling of energy from one of these fibre modes to the other one.

C3.1.5.10 Distributed sensing using a Sagnac loop interferometer

The use of Sagnac fibre-loop interferometers for detection of rotation has been discussed earlier. However, they can also be used to locate the longitudinal position of a time-varying disturbance acting at a non-central location on their sensing loops [123, 124].

The ability of this interferometer to locate the position of a time-varying disturbance relies on the counter-propagating nature of the light and the break in symmetry that the disturbance creates when perturbing the loop (see figure C3.1.42, but please initially disregard the branching into the delay loop and to detector 2).

When a time-varying strain, $\varepsilon(t)$, acts on a Sagnac loop of optical fibre, at a distance, z, from the centre of the length of fibre forming the loop, it perturbs the phase, $\phi(t)$, of the guided light. Due to the non-central location of the disturbance, light travelling in one direction is phase-modulated before light travelling in the other. This results in a net phase difference, $\Delta\phi(t, z)$, between the two returning counter-propagating wavetrains, when they return to the detector and interfere, at the output of the loop. It was shown by Dakin *et al* [123, 124] that, assuming a small slew rate, $\mathrm{d}\phi/\mathrm{d}t$, for the optical phase, then $\Delta\phi(t, z)$ is given by:

$$\Delta\phi(t, z) \propto \frac{2z}{V_\mathrm{g}} \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}$$

where V_g is the group velocity of the guided light.

This, however, presents a problem, as we have a sensor response dependent on two unknowns, firstly the rate of change of the phase perturbation and, secondly, its position, relative to the centre of the fibre forming the sensor loop. The initial solution, in the references above, was to separately measure the value of $\mathrm{d}\phi/\mathrm{d}t$. This was achieved using the delay loop in figure C3.1.42, that the reader was asked to initially disregard when describing the Sagnac loop response. This additional loop enables a fraction of the light which had travelled in one direction around the Sagnac loop to be mixed with a suitably delayed portion of light derived from the same original source. This 2nd delay loop was chosen to be of similar

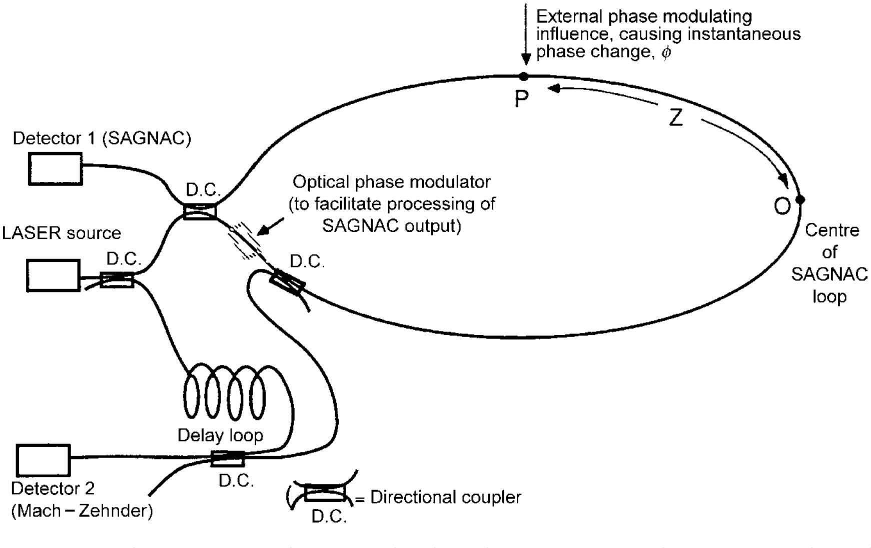

External phase modulating influence, causing instantaneous phase change, ϕ
P
Z
O
Centre of SAGNAC loop
Detector 1 (SAGNAC)
D.C.
LASER source
D.C.
Optical phase modulator (to facilitate processing of SAGNAC output)
D.C.
Delay loop
D.C.
Detector 2 (Mach – Zehnder)
D.C. = Directional coupler

Figure C3.1.42. First use of a Sagnac interferometer for disturbance location. The lower part shows how a separate delay loop can be added, to form a balanced Mach–Zehnder interferometer when light taken from the Sagnac loop is added to that from the delay loop.

length as the Sagnac loop to that point, so that a second interferometer, in this case a balanced-path fibre Mach–Zehnder could be formed, the output of which was detected to derive a measure of the actual value dependent on the phase change ϕ. Differentiation of the derived value of ϕ (or, if the phase changes are large and fast enough, simple detection of the frequency of fringe crossing by connecting the detector to a frequency counter or rate meter) yields a value for $d\phi/dt$. This new disturbance location method was applied to detect, and locate, fast thermal changes acting at different points of a 200 m length of fibre.

Following this first Sagnac disturbance location method, various new architectures using twin Sagnac configurations have been suggested, initially by Udd [125], followed by versions by Spammer *et al* [126–128], Ronnekleiv *et al* [129] and Fang [130]. All these avoided the need to accurately balance two independent optical paths. Such arrangements effectively used two Sagnac loops, configured to share a common fibre sensing section, but it was arranged, via several different optical routing means, that the optical path lengths of these two loops had different effective centres. These allowed two separate detector output values to be derived, now giving two response equations, and hence allowing calculation of the both of the two unknown quantities discussed earlier. The attraction of such configurations is that now the two counter-propagating paths in each Sagnac loop have the inherent advantage of such interferometers, in that they are both intrinsically path balanced. In addition, the reference point for the centre of the loop, where there is zero sensitivity, can conveniently be moved to lie outside the fibre sensing section if desired.

These new configurations were operated with either (wavelength multiplexed) twin-source configurations or with other more intrinsically lossy arrangements using directional 3 dB couplers and twin detectors. (The minimum theoretical loss of a dual-Sagnac system using 3 dB couplers is 18 dB in each Sagnac loop.) A recently reported method (figure C3.1.43) by Russell and Dakin [73] has improved designs further, and for the first time allowed the use of a single source and single detector, using appropriate WDM routing components to ‘slice’ the light from the source into two wavelength bands, and route each band of light around different Sagnac loops, again with different effective optical path centre positions, and the desired shared fibre sensing section.

Describing figure C3.1.43, a broadband (low coherence) Er^{3+}-doped fibre superluminescent source was spectrally sliced into two wavelength bands using wavelength-division multiplexers (WDMs). These

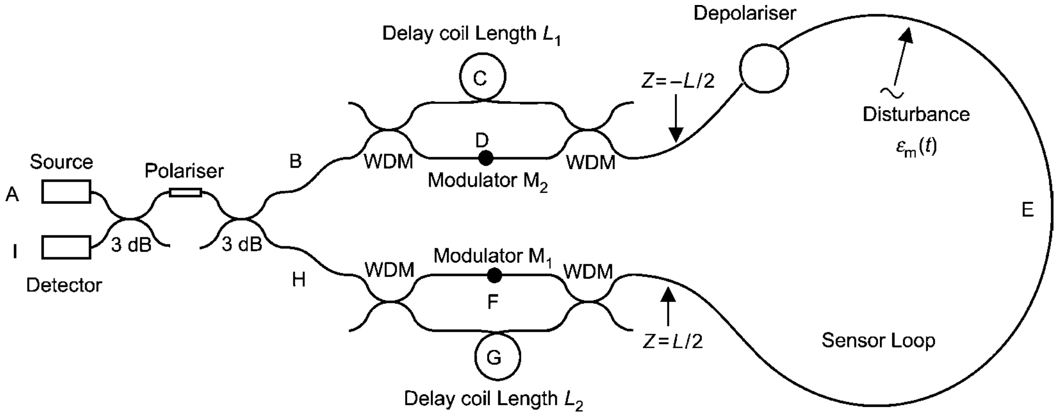


Figure C3.1.43. Twin-Sagnac-loop disturbance location system, using ‘spectral slicing’ of a broadband rare-earth-doped fibre light source to obtain two wavelength channels, and WDM routing components to construct two optical loops with effectively different centres. Using modulation signals of different frequencies, at M_1 and M_2, the beat signals on one detector can be separated electronically.

routed the light along two, essentially independent, Sagnac interferometer loops, one for each wavelength. The first Sagnac loop was defined by the bi-directional path ABCEFHI, read clockwise around the sensor. This circuit includes a fibre delay coil, C, a sensing fibre length, E, and a piezoelectric phase modulator, F. Similarly, the second Sagnac interferometer was defined by routing light bi-directionally around a path labelled ABDEGHI. This again includes a piezoelectric modulator, D, a common sensing fibre length, E, and a fibre delay loop, G. The effective centres of each of the sensor loops are offset (in opposite directions) by half of the path length in the delay coils labelled C and G. This ensured that each Sagnac gives a different response to a common perturbation, despite sharing a common fibre section, allowing simultaneous evaluation of the position of the disturbance, its amplitude and its rate of change.

Each Sagnac was phase-biased [131] with sinusoidal strain signals of different frequencies (f_1 and f_2), each being one of a set of natural eigenfrequencies of the fibre loop. The bias frequencies (f_1 and f_2) were chosen such that the magnitude of their difference frequency $|f_1 - f_2|$ was above the frequency range of the expected disturbance signals (i.e. the base-bandwidth of the output of the sensor). This phase bias allows both of the interferometers to share a common optical detector, as it provides amplitude-modulated carriers of a different frequency for each signal generated by each Sagnac.

This system used data-acquisition hardware to sample the sensor outputs in real-time and a software system which could be instructed to either (a) locate the three largest disturbances observed in the frequency domain or (b) locate a disturbance of selected character, i.e. of a specific frequency or amplitude. The positions and the frequencies of each of the disturbances were then calculated and displayed. Using the system, it was possible to monitor sinusoidally varying phase disturbances of only 0.025 radian phase change amplitude (~1.3° optical phase change, or equivalent to a few nm fibre stretching) acting on a 40 km long sensing loop and locate them with 100 m position resolution.

C3.1.5.11 Distributed sensing using a counter-propagating optical pump pulse

If an optical signal from a steady-state, or CW source is transmitted through a fibre to a detection system, the power level received will be dependent on the total attenuation in the fibre. If, however, an intense optical pulse is now launched into the optical fibre, in the opposite direction (see figure C3.1.44), the intensity of the transmitted CW lightwave will now be affected by any optical gain processes. Such effects can arise from several possible nonlinear interactions with the pump.

The first report of such a system was presented by Farries and Rogers [132]. They used a pulse, from a NdYAG-pumped dye laser at 617 nm, to provide Raman gain in a continuous length of monomode fibre. The CW beam was a 633 nm signal from a helium–neon laser source. The Raman gain is very sensitive to polarization and therefore the arrangement is capable of detection and location of lateral stresses in a fibre of low intrinsic birefringence, as these cause polarization mode conversion, hence modifying the degree of Raman gain.

The technique was a major advance, as it was the first of a class of sensors, but suffered, in its early form, from a number of practical disadvantages. As often the case for a first laboratory form, it used rather large, and hence inconvenient, laser sources, and it is likely the results would have been critically dependent on the pump power level of the dye laser source. In addition, it would also be likely to suffer badly from undesirable polarization and other variations due to environmental factors such as bends in the fibre.

More recently, sensors using stimulated Brillouin scattering have been reported, one of the most significant being the Brillouin optical time-domain analysis (BOTDA) system devised by Horiguchi [109, 110], which has since formed the basis of a number of systems for sensing strain in very long fibres, using more complex versions of the simple arrangement of figure C3.1.44. This sensor takes advantage of gain from the stimulated Brillouin scattering process. As with the spontaneous Brillouin scattering process

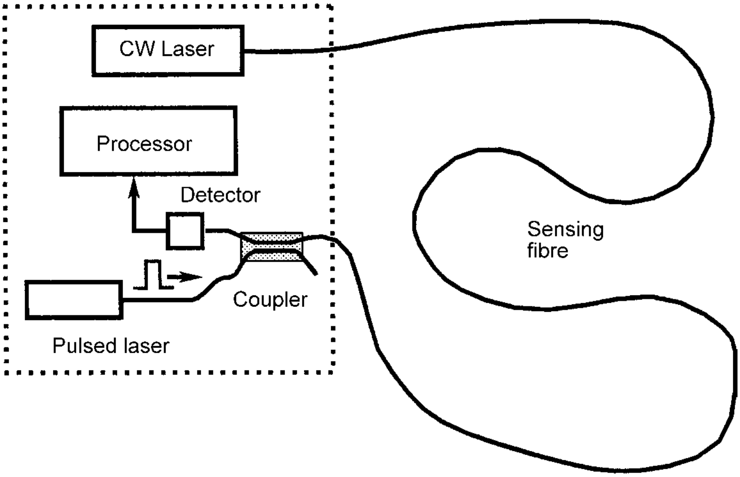


Figure C3.1.44. General concept of a distributed sensor using counter-propagating optical waves. The high intensity pump signal causes nonlinear optical gain processes to occur, which changes the intensity of the CW probe signal with time.

mentioned earlier, the centre line of the Brillouin gain curve is offset from the pump laser signal. Using a separate tuneable laser (or more conveniently by deriving a frequency-shifted signal from the one laser source) pump and probe beams can be arranged to be suitably offset in frequency, to not only ensure the desired Brillouin gain occurs, but also, by frequency sweeping, actually measure the frequency of peak gain. Taking advantage of the time-dependent nature of the signals when the pump is a short pulse, the distribution of Brillouin gain shift versus distance can be derived by a combination of sweeping the laser frequency offset and observing the temporal patterns.

C3.1.5.12 Use of the 'synthesis of the optical coherence function' method for distributed sensing by monitoring the gain process associated with stimulated Brillouin scattering. (K Hotate, RCAST, University of Tokyo)

Distributed strain sensing based on Brillouin scattering, as discussed earlier, is a promising technique for 'smart materials' and 'smart structures' applications, and sensor systems using Brillouin OTDR are already on the market. However, the technique developed so far has a spatial resolution limit of several metres. In smart structures applications, for example, in an aircraft's wing, this limit would be a major disadvantage. The resolution limit of the conventional technique is a consequence of its pulse-based nature. An optical pulse for generating Brillouin gain has to be longer than the damping time of the acoustic wave. With shorter pulses, the Brillouin gain spectrum (BGS) broadens out, making it more difficult to determine precisely the frequency of the spectral peak. Since the typical value of the damping time is 25 ns, the practical limit for spatial resolution turns out to be typically about 1 m.

To circumvent the resolution limit, a new technique has been developed [120]. It is based on the control of the interference between the pump and probe lightwaves that excite stimulated Brillouin scattering (SBS). SBS requires interference between two counter-propagating lightwaves. In the technique, we control their coherence, so as to localize the SBS at a specific position in an optical fibre, where their correlation is high.

Figure C3.1.45 shows the proposed system for measuring the distribution of the BGS along an optical fibre. The light from a 1.55 μm frequency tunable DFB-LD is split by a coupler, to provide light

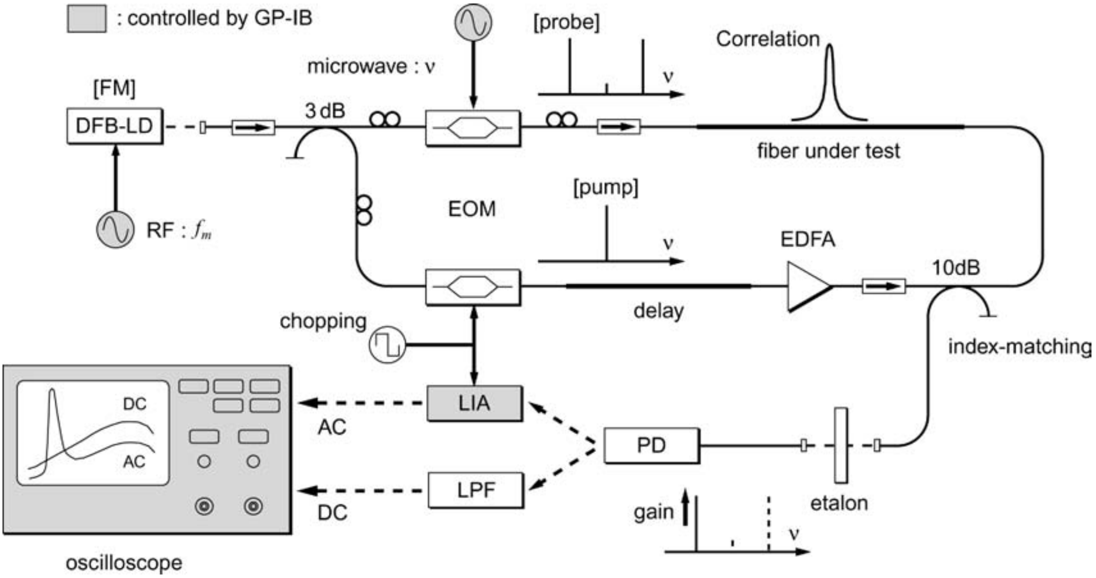


Figure C3.1.45. Schematic of system for high resolution sensing, using stimulated Brillouin backscatter in conjunction with the 'synthesis of optical coherence function' method of K Hotate.

sources for the pump and probe lightwaves. Light from one output of the coupler is intensity modulated (chopped) at a radio frequency, using a $LiNbO_3$ electro-optic modulator (EOM), with an electrical drive signal. The output is amplified by an erbium-doped-fibre amplifier, and then launched into the fibre under test to serve as the optical pump (this is shifted in frequency by only a very small amount by its intensity modulation). The other output is intensity modulated, at a much higher microwave frequency, ν, by a second $LiNbO_3$ modulator, so that AM sidebands are generated around the incident lightwave of a frequency ν_0. The lower sideband, at $\nu_0 - \nu$, is used to serve as the probe beam and the unwanted residual sideband and carrier are eliminated by an optical filter. This probe beam propagates in the fibre under test, in the opposite direction to the pump, until it reaches the detector.

An important point is that the pump and the probe undergo identical frequency modulation, as they both arose from the laser diode source, and the system takes advantage of its capability to be frequency modulated directly, by changing its bias current. As a result, SBS occurs exclusively at the position of peak correlation, the only point where the two lightwaves are well correlated. The correlation peak can be conveniently shifted along the fibre, simply by changing the FM frequency f_m of the laser diode source. The increase in the probe power, resulting from Brillouin gain in the fibre, is detected synchronously using a lock-in amplifier. This was referenced using the electrical drive signal used to chop the pump intensity. We obtain the BGS by varying the frequency of the microwave signal used to generate the frequency shift in the probe signal. By repeating the BGS measurement over the appropriate range of different frequency drives, f_m, to the FM, the BGS is obtained as a function of position along the fibre.

Application of the correlation-based Brillouin sensor has been investigated [120] for the measurement of strain distribution in small-scale (few cm) material samples. The sample material, chosen to demonstrate this in the laboratory, was a cylindrical acrylate-ring coil former, which could be stressed by side pressure to deform it. A single sensing coil of dispersion-shifted fibre was first wound around it, then bonded to it with epoxy cement (figure C3.1.46(*a*)). The coil former could be deformed to an approximately oval cross-section to test the sensing system.

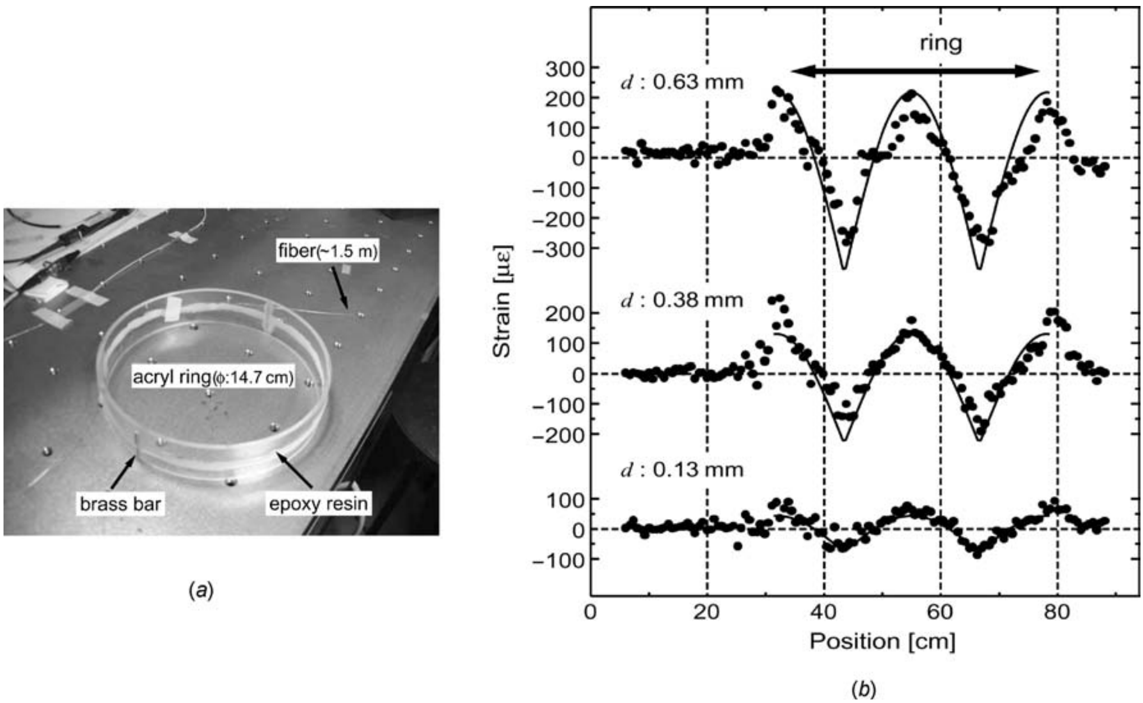


Figure C3.1.46. (*a*) Photograph of an acrylate ring sample with a single turn of fibre bonded to it. (*b*) The measured strain patterns using stimulated Brillouin backscatter in conjunction with the 'synthesis of optical coherence function' method.

The frequency shift of the modulator was set to 3.2 GHz, corresponding to a 1 cm spatial resolution and the strain distributions for various levels, *d*, of ring deformations were measured. Figure C3.1.46(*b*) shows that, for small ring deformations, the experimental results agreed well with the theoretical value (solid lines), but deviation from the theoretically expected value became greater at high deformation values. This is because the spatial resolution became insufficient to accurately trace out the step strain gradient at high deformation values. In this scheme [120] a reasonably high measurement sampling rate, of several tens of Hz, was also been realized.

C3.1.5.13 Distributed sensing to determine profiles of strain and temperature in long in-fibre Bragg gratings

As a final note on distributed sensing, it should be noted that there are several methods capable of resolving the effective peak reflective wavelength along a fibre Bragg grating, as a continuous function of position. Methods to produce very long fibre Bragg gratings have been discovered some years ago [133] and gratings of up to 1 m in length are now almost routinely manufactured, primarily intended for dispersion compensation in optical telecommunications systems. Methods to determine the variation of the peak reflective wavelength along the length of such gratings are available, making them highly attractive starting components for distributed sensors to measure in-fibre strain and/or temperature. Very high spatial resolution (typically 0.1 mm) over distances up to 1 m is possible, allowing of the order of 10^4 sensing elements! The methods devised to address such gratings are now quite numerous and some are fairly complex, so only a brief outline will be given here.

Early methods made assumptions about the nature of the index modulation in the grating, in order to derive optical delay (two-way time of flight) information from the wavelength components reflected [134] or to get wavelength from the delay [135]. Various means, such as tunable lasers have been used to determine the wavelength information, and interferometric phase determination has been used to detect the time delay.

One attractive method of defining, more precisely, the distance to the measurement point in the grating is to use low coherence interferometry. This operates in a similar way to that of coherence multiplexing and to the distributed sensing methods described above. Volanthen *et al* [136, 137] used this method, where the long Bragg grating to be interrogated (initially a chirped one, 50 mm long was used) was in one arm of a Michelson interferometer, whilst the other arm contained a fixed-wavelength much-shorter Bragg grating. The effective distance to the grating in this second fibre arm could be changed in length, using a fibre stretcher, to determine the position of good fringe visibility in the long Bragg grating in the other arm. The attraction of this method was that no prior knowledge of the grating wavelength versus distance profile was needed, and, unlike the earlier ones, the method could work with gratings that did not have a monotonic variation of wavelength with distance.

Some recent systems have used commercial coherence domain reflectometers, which conveniently contain an in-built swept Michelson interferometer, to determine the point in the long grating [138, 139]. This grating can then be monitored at any point, at any one moment in time, with an external acousto-optic tunable filter, which allows the scanning of interrogation wavelength. In these systems, an external optical amplifier was used to compensate for fibre system losses by boosting the signal strength from the ELED source in the instrument. Such sensors may have useful applications for monitoring strain distributions in complex mechanical components during factory testing, but are likely to remain rather costly and complex for many vehicular applications, such as aerospace monitoring.

In the final section of this chapter, we shall now move on from discussion of many sensor concepts to present a more specific case study of how one type of sensor has addressed several industrial metrology problems.

C3.1.6 High resolution measurements using low-coherence interferometry: a practical industrial case study

This last section presents an industrial sensor case study, to illustrate how an interferometric sensor has been designed, constructed and used in real production applications.

C3.1.6.1 Introduction to instrument design

The sensors described here are all variants of an interferometric sensor, based on a dual Michelson arrangement, having sub-micron distance accuracy for measuring optical distance. Several versions have been developed, for use in a variety of in-factory quality control and process monitoring applications. The basic system [140] includes both coherent and low-coherence light sources and employs the well-known optical configuration of the autocorrelator (see figure C3.1.47).

Light from the low-coherence light source is guided to the test sample via a single-mode fibre. Light beam components, which are partially reflected from each of the optical interfaces in the sample, mix in the interferometer section, to give a fringe pattern that is dependent on the type of source used and on the various optical path differences. The detected response, as the interrogating interferometer is scanned in length, is used to determine sample-dependent parameters, such as optical distances. When reflection occurs from input and exit surfaces, the optical thickness of sample can also be measured.

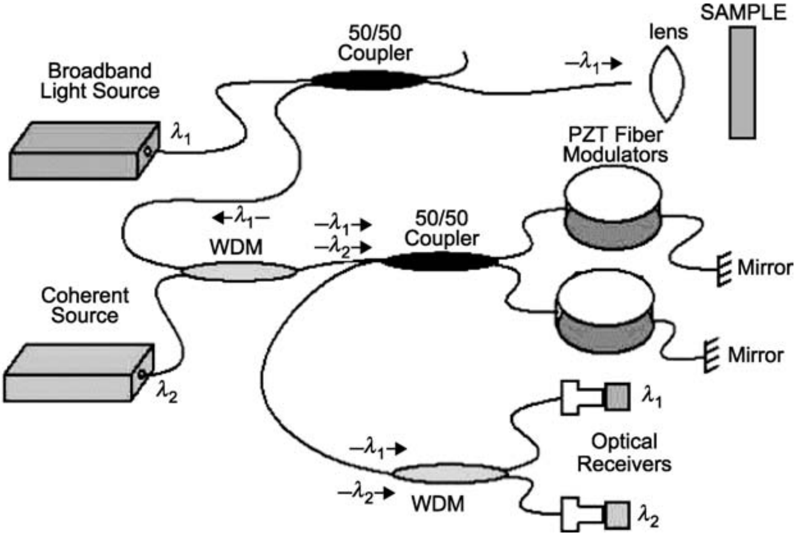


Figure C3.1.47. Diagram showing the experimental arrangement of the all-fibre interferometer. The mirrors shown are Faraday mirrors, to prevent polarization fading.

The different light paths of the broadband light, when travelling from source to detector, form what is commonly called a 'white-light interferometer' [141]. This has the feature of only being capable of seeing visible fringe patterns when the interfering optical paths are very closely matched.

The narrow linewidth laser source always creates visible fringes, as its coherence length is greater than all of the optical paths involved, so use of this second light source allows fringe counting to be used over all distance ranges, and permits extremely precise monitoring of the optical distance through which the interrogator interferometer has been moved. This count is also used to trigger, at certain user-determined fringe-count numbers, data acquisition of the interferometric data from the broadband source, ensuring this is taken at constant optical distance intervals. The use of the coherent laser source therefore ensures that all the parameters calculated from the detected signals have a 'built-in' optical distance scale, calibrated to the narrowband laser wavelength.

Figure C3.1.47 shows a schematic of an all-fibre version of the measuring interferometer. Light from a short coherence length light source (typically a 1310 nm ELED, of ~55 nm bandwidth) is passed through a 2 × 1 coupler (50/50 combiner/splitter) and is directed through a lens, before being incident on the sample under test. Light returning from the sample passes through the coupler and into a WDM (wavelength-selective) coupler, which is used to combine it with other light from a highly coherent laser source. A fibre Bragg grating-stabilized diode laser, at ~1550 nm, having less than 5 ppm frequency drift, is presently utilized as the coherent light source. The combined light passes through a second 2 × 2 coupler directing light into a pair of matched-path-length fibre coils, which are wrapped around piezoelectric PZT cylinders. These are used to stretch the fibres and change the path lengths of the two arms of the Michelson interferometer. In order to prevent 'polarization fading', Faraday mirrors are placed at the ends of the fibres containing the coils. These have the property of reflecting the light back through the coils in such a manner that they will eventually leave the recombining coupler and impinge on the detector with always the correct polarizations to interfere.

The light components from the two arms of the interferometer interfere as they pass back through the 2 × 2 coupler and pass through a WDM coupler, which separates the 1300 and 1550 nm light regions. This separated light goes to a pair of detectors. The zero crossings of the detected sinusoidal signal from the 1550 nm laser are used to trigger data acquisition of the low coherence signal. Because of the two-way propagation of the light, the distances between zero-crossing points represent optical path length changes in the measurement region equivalent to only 1/4 of the monitor laser wavelength. Data acquisition is performed utilizing a National Instruments PCI-6111 5 MHz, 12-bit A–D converter board. A 2 GHz Pentium IV computer, operating in a 'LabView 6.1' environment, with a Windows 2000 operating system, is used as the signal processor. A measurement rate up to 10 000 thickness measurements per second has been demonstrated, utilizing sampling intervals corresponding to a path length of $\lambda/4$ of the laser. The PZT cylinders are operated in a push–pull manner, so that the optical path length in one fibre increases, as the path length of the other fibre decreases.

Application-specific components of the system include the optical probes, optical switches, sample and probe transport mechanisms and the peak processing algorithms used to provide sample information and user interfaces. The equipment has been developed for measurement of thickness, and thickness profiles, of polymer films [142], measurement of optical retardation in films [143], measurement of liquid thickness distributions on coating hoppers [144], the length calibration of optical cells, assessment of the focus of digital camera imagers [145, 146] and surface profile measurements of films, wafers and imagers. There are clearly many other potential uses of this versatile telemetry method.

For the low-coherence light source, interference with visible fringes only occurs when the path lengths of the two arms in the interferometer are equal to within a few coherence lengths. In order for any interference to occur, light must of course be reflected back into the interferometer from the sample. This will occur due to Fresnel reflection at each optical interface in the sample. The distance between adjacent interference peaks is a measure of the optical thickness (group index of refraction, n, times the true physical thickness) of the sample material. In air layers, the distance between the two adjacent surfaces is approximately the thickness of the layer, as n_{air} is very close to unity. Since the instrument uses a stabilized laser light source to provide constant distance interval measurements, the instrument measures *optical* path distance, defined, in regions of media other than vacuum, as n times the physical thickness.

As mentioned above, the measurement configuration of the interferometer is that of the optical *autocorrelation* mode, in which light reflecting from the sample is input to both arms of the Michelson interferometer. In the autocorrelation mode, light beams reflecting from the sample are made to interfere in a way that both arms of the interferometer see reflections from all of the optical interfaces in the sample (as an example the front and back surfaces of a film). As the path lengths of the coiled fibres in the interferometer are changed, a series of visible clusters of interference peaks are observed, indicating the optical path differences between adjacent optical interfaces. The zero-path-difference, or self-correlation condition occurs at the point when the two path lengths of the Michelson interferometer become equal; in which case, all optical interfaces in the sample give strong visible interference fringes. The measured distance between the largest peak, corresponding to this zero path length difference, and the first set of adjacent peaks, is a measure of the shortest optical path difference in the sample. One major advantage of using the autocorrelation configuration is that the interferometer can be located at any distance away from the sample interface, provided the fibre losses permit it. A system with the measurement head located as far as 10 km from the interferometer has been demonstrated.

The system processor uses a peak-location analysis technique, to find the true centre of the envelope of a cluster of interference fringes. This envelope is, to a first approximation, a cosine function, having a Gaussian intensity-modulation envelope. The peak location algorithms that have been developed [146] include moment calculations, Gaussian-peak analysis and Fourier-phase-slope analysis [147]. When using a sampling distance interval of $\lambda/4$, they provide measurement repeatability better than 0.050 μm (50 nm), with samples having interface separations greater than 20 μm. Once the peak locations are calculated, the appropriate distances relevant to the measurements being performed must be computed. Distance calculations are noted and compared to defined acceptance ranges and thresholds.

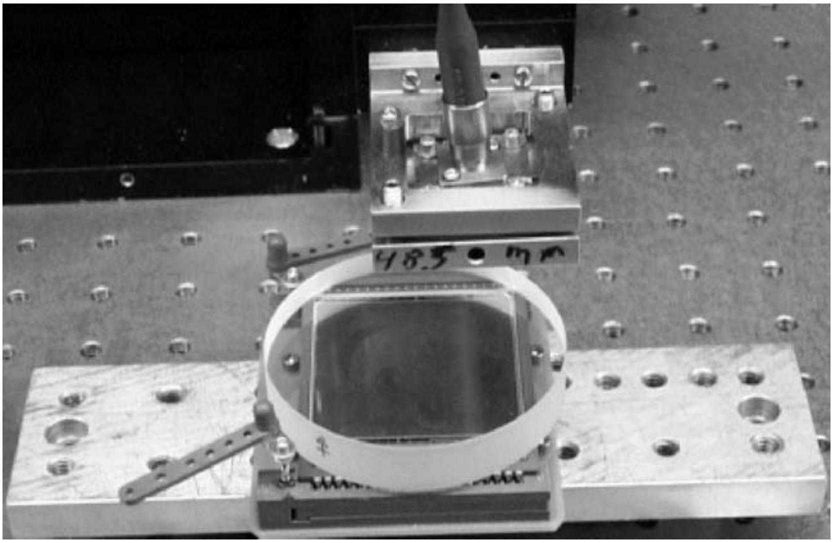


Figure C3.1.48. Photograph of testing of early optical arrangement for the probe, showing it clamped to a camera imager detector array, in order to check the distance to the detector surface.

C3.1.6.2 Design of a robust sample interface probe

In order to apply the instrument to an industrial process, a robust optical interface had to be developed. An early trial of the optical arrangement is shown in figure C3.1.48, in this case with the system being tested with the optics clamped to an imager chip.

An angled fibre optic probe, designed to eliminate back reflections from the fibre connector, has been used in the probe head. The fibre/lens coupling arrangement used in all these probes is shown at the top of figure C3.1.49. An APC-FC single-mode optical fibre is used, with an APC-FC side attached to an angled FC receptacle mounted onto a two-axis 'gimbal' mount. The APC-FC side has an 8° cleave angle (the angle is exaggerated in the diagram to emphasize it is a non-normal surface) to effectively eliminate any back reflection from the fibre end coupled to the focusing GRIN lens. The APC-FC receptacle is attached to a mount holding the GRIN lens, with a bracket having a 3.7° offset angle, so that, by tilting the optical fibre coupling, the output light can be arranged, after refraction, to exit normal to the plane of the GRIN lens, as the system requires. The entire assembly is mounted in a two-axis 'gimbal' mount, which allows the angles to be varied independently, with respect to the x- and y-axes. An optional micrometer stage, for adjusting z height, can also be utilized, if desired.

The lower part of figure C3.1.49 shows examples of more highly developed and more robust quick-connect designs, more suitable for camera testing.

These designs use a fibre with an APC-FC ferrule on one end of the fibre cable, mounted centrally in a custom chuck, allowing the light to exit parallel to the axis of the chuck. The lens is inserted into a ridge from the other end of the chuck, so that it will be at a known distance from the end of the APC-FC ferrule. The distance between the ferrule tip and the lens determines the focal length of the optical probe.

C3.1.6.3 Assessment of in-camera optical focus distances

We will now present, as a detailed case study of fibre optic-based interferometry, an application for focus verification in professional digital cameras. As any amateur user can easily see when changing the film in

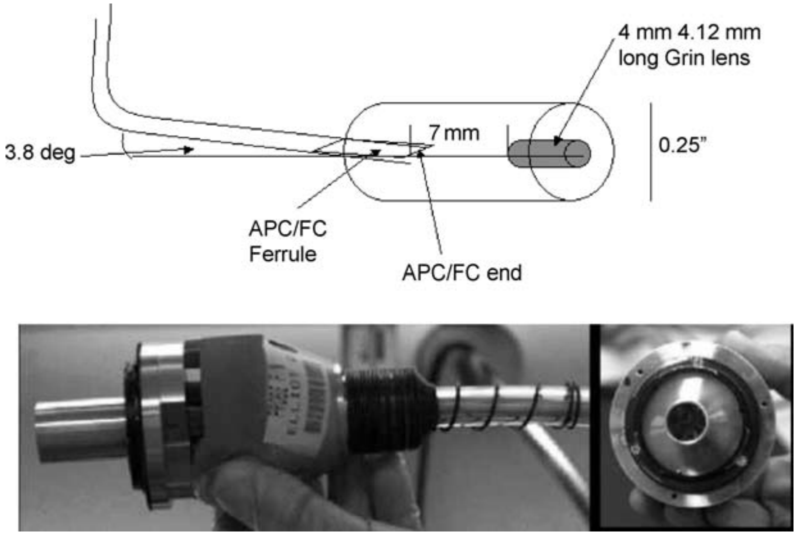


Figure C3.1.49. Design and construction of probes for determining in-camera imager positions. Upper curve shows the angled fibre design and the GRIN focussing lens, and lower photographs show the probes with standard camera fittings.

a 35 mm SLR camera, the film normally rides on parallel metal rails, which determine the position of the desired focal (film) plane in the camera. In digital cameras, there is of course no film; instead a semiconductor chip imager detector is used, usually in a hermetically sealed package, with the detector chip behind a glass window. Often, the detector does not have a precisely known position in the package, so simple external location, using the previous method of film rails, or other form of mechanical end-stop, would lead to uncertainty in the position of the detector surface. Thus, an accurate and repeatable method is needed to determine the depth of the front surface of the imager detector, relative to the lens flange mount. For good focus, position tolerances are required to be $\pm 25\,\mu\text{m}$, or less, over the full surface of the imager. The measurement must be performed rapidly, in order to be practical for a camera production line. Since non-scientifically trained operators on the assembly floor will perform the measurement, it must be robust, easy to operate, and allow camera quality certification to be based on the measurement results.

The optical probes shown in the bottom of figure C3.1.49 were developed to be inserted into a camera body by a production line operator, in order to certify that the imager was installed at the appropriate focal plane for a given camera model. The probes include a thick optical flat near the end of the cylinder, which was designed to act as a reference surface when performing the in-camera measurements of location of the imager surface. This reference surface has to be within the focal distance of the optical probe, as does the relevant surface of the imager being tested. At the camera-mounting end, a standard lens-flange mounting ring is also included, to allow mating to the camera in the same manner as, but instead of, a standard lens. The optical probes usually include multiple fibre chucks each at different positions, so the centre and the corners of the imager can be tested, to determine any undesirable tilt of the imager plane in the camera. In order to enable multiple point measurement, an optical polling switch is placed in-line with the fibre coming from the instrument. Sequential measurements, first at the centre, and then at the four corners of the imager, can be performed during a measurement.

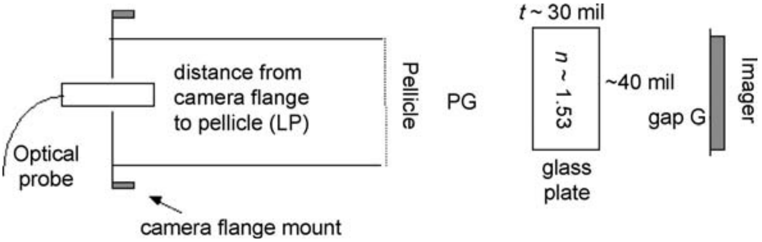


Figure C3.1.50. Schematic showing the measurement geometry for in-camera measurements.

A schematic of the measurement geometry is shown in figure C3.1.50. For a particular camera model, the imager focal depth should lie at an optical depth equivalent to a distance F_0 (mm) in air. During the manufacturing of digital imagers, a glass plate, of thickness t and index of refraction n, is mounted in front of the imager, with a gap G between the imager and the glass plate. During manufacturing, the imager and glass plate are packaged together and hermetically sealed into a single unit. When a flat glass plate, of thickness t and index of refraction n, is inserted into the optical path of a lens, the effective focal length of the lens shifts by an amount $\Delta_g = t(1 - 1/n)$. Hence, the required depth of the imager plane in a camera body changes as the thickness of the glass cover plate is varied. Hence, the correct focal position for the imager plane with a glass cover is now given by $F_0 + \Delta_g$. Figure C3.1.50 shows the measurement relationships used to determine focus error from the measurement geometry.

The purpose of the in-camera measurements is to determine the focus error (defined as the distance between the actual focus position and the ideal focus position) at each of the measurement points over the surface of the imager. The focus error is calculated from the expression

$$\text{Focus error} = \text{LP} + \text{PG} + t + G - t(1 - 1/n) - F_0$$

where LP is the distance from the lens flange mounting ring to the optical probe reference surface, PG is the distance between the probe reference surface and the top of the imager cover glass, t is the thickness of the imager cover glass, n is its index of refraction and G is the gap between the imager surface and the bottom of the cover glass. Calibration for in-camera measurements used a reference cradle, which had a lens flange mounting ring to receive the optical probe and a flat plane located a known distance away from the lens flange mounting ring plane, which was used to measure the distance LP.

Figure C3.1.51 shows an example interferogram, with an indication of the origin of each of the groups of interference peaks. The definitions are as follows: PG is the distance from the pellicle to the imager cover glass, nt is the imager cover glass optical path (n = glass group index of refraction and t = glass thickness), and the gap, G is the distance from the imager cover glass to the imager plane. When light passes through the optical probe shown in figure C3.1.50, reflections occur at the pellicle, at the front of the imager cover glass plate, at the back of the imager cover glass plate and at the imager surface. Reflected light from each of these interfaces is input into the interferometer. Constructive interference occurs when the differences in path lengths of the two arms of the interferometer are equal to distances between adjacent and non-adjacent peaks.

During a measurement sequence, the interferometer was set up to scan, symmetrically, over a distance sufficient to observe all the peaks of interest. For the in-camera focus assessment application, the all-fibre interferometer was typically scanned, over a total scan distance of about 6 mm, centred around the zero crossing distance, using a scanning-motor frequency of 50 Hz. In the symmetric mode, four key measurements were taken in each scanning-motor cycle to cover all the parameters of interest. Measurements were typically made for a 0.1 s duration. The large amplitude, zero-crossing peak, is at the centre of the scan. A threshold level was set for all peaks in the interferometer and the locations of the zero

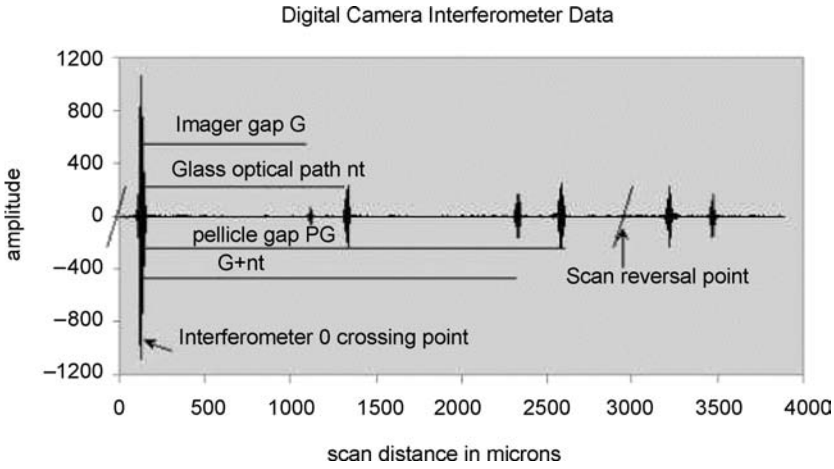


Figure C3.1.51. Typical measured interferogram, when interrogating a digital camera.

crossing peaks were determined. The distances between all the other peaks in the interferometer trace and the zero crossings were then calculated, with respect to the nearest zero crossing peaks. All peaks of interest were placed into acceptance ranges, and the quality assurance statistics were then calculated.

The appropriate acceptance ranges for the parameters of interest for the in-camera focus assessment application for a particular camera are as follows. The imager gap is specified to be in the range of 800–1100 μm, the optical thickness of the cover glass *nt* is in the range of 1100–1250 μm, and PG is in the range of 2500–2800 μm. The index of refraction of the glass is known to be 1.5335. The focus error should be less than ±50 μm.

Figure C3.1.52 shows a camera being brought to the in-camera measurement fixture. The operator first opens the camera shutter and then inserts the camera into the lens flange-mounting ring of the optical probe. The operator then reads the bar-code on the catalogue tag number of the camera, so it is displayed in the appropriate 'new part number' box on the measurement screen shown in figure C3.1.53.

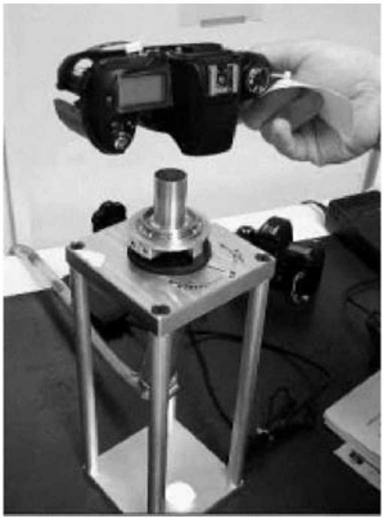
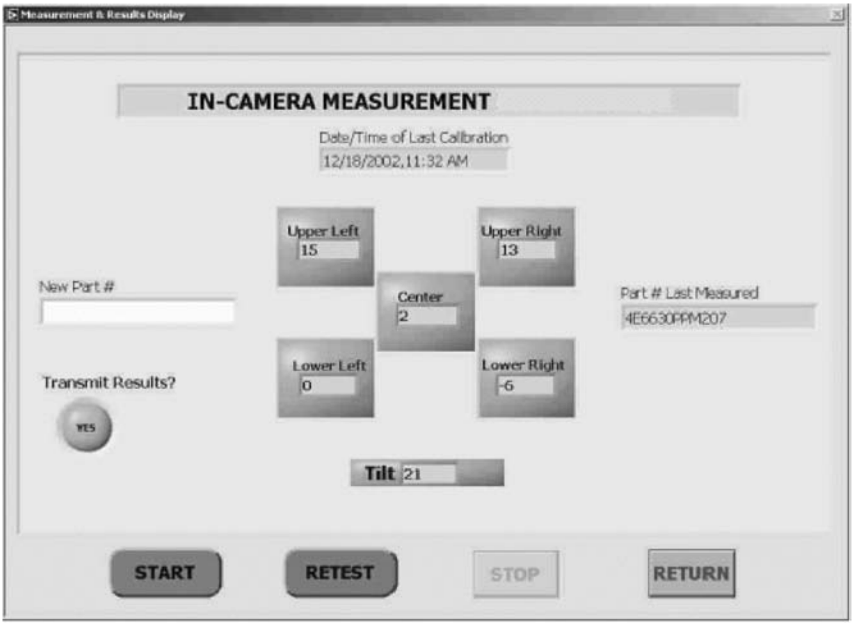


Figure C3.1.52. Test assembly with a camera being brought for testing.



Date/Time of Last Calibration: 12/18/2002,11:32 AM
Upper Left: 15
Upper Right: 13
Center: 2
Lower Left: 0
Lower Right: -6
New Part #:
Part # Last Measured: 4E6630PPM207
Transmit Results?
Tilt: 21

Figure C3.1.53. The display of camera test data on a PC, as used for quality assurance tests.

The start button is then clicked on with the mouse and the measurement sequence starts. The probes are sequenced in the order centre, upper left, upper right, lower left and lower right. After 0.5 s all the data are shown. Green lights indicate that the numbers are within specifications and red lights are used to indicate a problem. When the transmit results button is on, the results are automatically transmitted to a network database. The screen also shows the last calibration and the previous measurement result.

Many other types of measurements have also been developed by Kodak, using the same concept, including a film-rail position assessment and a die to plate assessment. Clearly, many more industrial metrology applications are possible for such devices. One application of particular general interest is surface profiling and film thickness measurement, which will now be briefly discussed. Finally, a photograph of the same method used for measuring liquid thickness in a coating plant will be shown.

C3.1.6.4 Surface profiling and extension of concept for simultaneous combination of thickness, surface and refractive index profiling (SSTIP)

Surface profiling is an important aspect of many industrial processes. We shall first describe its application for the measurement of flatness of large imagers and wafers. Here, a large area thick optical flat is placed above an imager that has been glued into a package or a large wafer. It is coupled to an *XY* scanning frame.

Figure C3.1.54 shows the surface profile results, at a measurement rate of 200 Hz, with the instrument set to take y-axis steps of 0.25 mm, as it is scanned at a rate of 50 mm s^{-1} along the x-axis. The data are shown relative to the best reference plane and are inverted. It can be seen that this imager has a slightly convex surface, being bowed upwards by about 8 μm in the centre. Profiles can also be obtained before, during and after curing of the imager into the package.

We shall now discuss how we have extended the concept, to perform very rapid profiling of several simultaneous properties of thin films, as they pass rapidly through the measurement rig. Figure C3.1.55

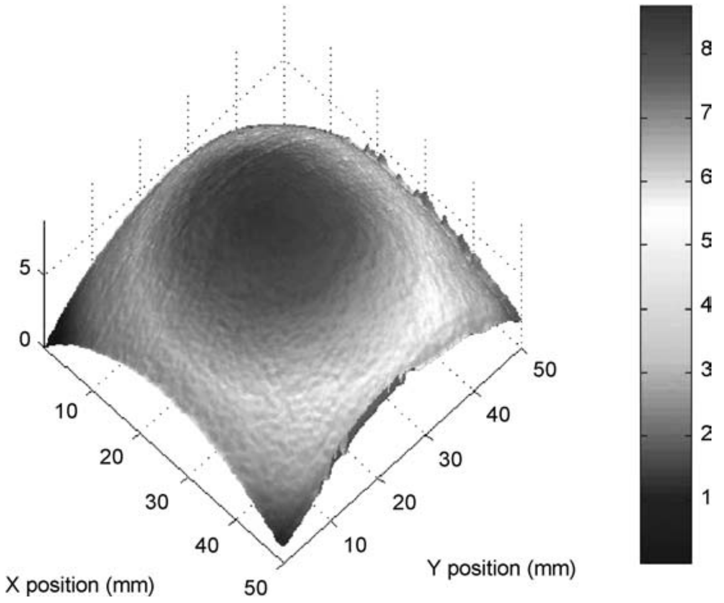
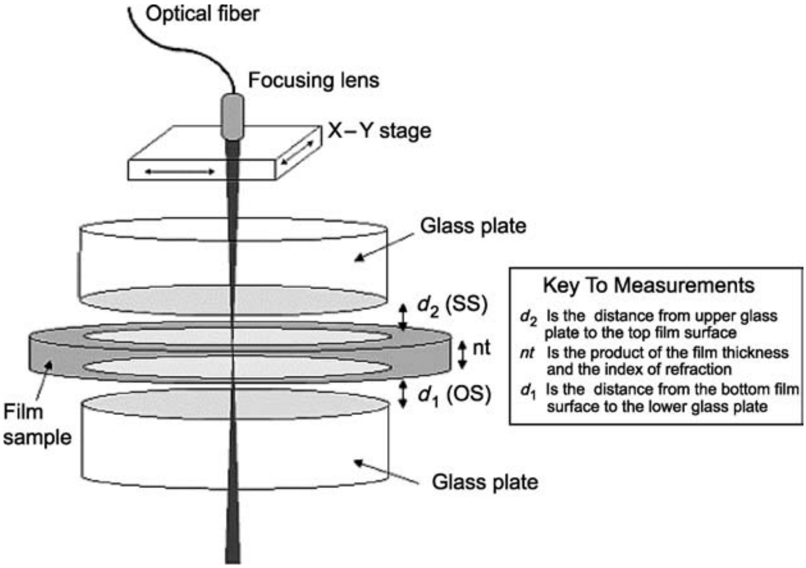


Figure C3.1.54. Colour-coded surface profile scan of an imager chip, showing that the imager surface is slightly convex (approximately 8 μm higher in centre).

shows an optical cell design used for simultaneous thickness, surface and (refractive) index profiling (SSTIP) of such thin films.

In this case, a film sample is placed between a pair of thick optical flats. Figure C3.1.56 shows part of a sample SSTIP interferogram trace.



Optical fiber
Focusing lens
X–Y stage
Glass plate
d_2 (SS)
nt
d_1 (OS)
Film sample
Glass plate
Key To Measurements
d_2 Is the distance from upper glass plate to the top film surface
nt Is the product of the film thickness and the index of refraction
d_1 Is the distance from the bottom film surface to the lower glass plate

Figure C3.1.55. System for obtaining simultaneous measurements of surface profile, film thickness and film refractive index (SSTIP).

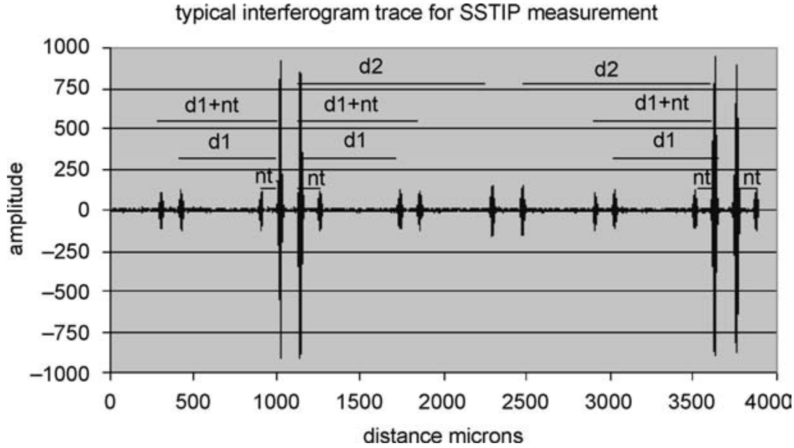


Figure C3.1.56. Interferogram when taking SSTIP measurements.

The largest peaks in the interferogram are the 4th, 5th, 14th and 15th from the left side of figure C3.1.56. The largest peaks are zero-crossing points, at which the two arms of the interferometer are of equal path length. The interferometer is deliberately set up with an offset in the zero-crossing position. The vertical dashed lines, between peaks 4 and 5, peaks 9 and 10 and peaks 14 and 15, indicate the positions at which the direction of interferometer scanning changes. Distances from the zero crossings to the smaller peaks are indicative of the distances between the various optical interfaces. The definitions for the various distances include film thickness, t, film group index of refraction, n, the distance from the bottom optical flat and the bottom surface of the film, d_1, and the distance from the top optical flat to the top surface of the film, d_2. In order to obtain an interferogram with the peak order shown in figure C3.1.56, the following relationship must hold true for the various distances between the measurement surfaces:

$$nt < d_1 < d_1 + nt < d_2 < 2d_1.$$

It is important that the optical flats should be nearly parallel and that the above relationships should hold over the entire surface of the film to be measured. When two of these distances happen to lie within 25 μm of each other, the interferometer may be unable to distinguish which surfaces in the sample the various peaks arise from. If the relationships between various distances changed significantly over the surface of the sample, the present software would not be able to keep track of which surface is related to which observed interferogram peak. Presently, we set up acceptance ranges for these measurements and keep the known order of peak distances over the entire surface. In the data shown in figure C3.1.56, the parameter values are:

$$nt = 121.2\,\mu\text{m}, \quad d_1 = 621.4\,\mu\text{m}, \quad d_2 = 1098.7\,\mu\text{m}.$$

The acceptance ranges used for the measurement were:

$$80 < nt < 200$$

$$580 < d_1 < 680$$

$$1050 < d_2 < 1150.$$

In order to calculate the parameters of interest, it is necessary to know the distance, d_0, between the optical flats. To do this, d_0 is first measured with the interferometer system, without the film being present. When the film is scanned through, the distance, d_1, measures the film's bottom surface profile and the distance, d_2, measures the top surface profile. The following relationships are used to calculate the index of refraction, n, and thickness t at locations, x, y, in the film sample:

$$t(x,y) = d_0(x,y) - d_1(x,y) - d_2(x,y)$$

$$n(x,y) = nt(x,y)/t(x,y).$$

C3.1.6.5 On-line thickness measurements

We shall very briefly show a photograph of one final system, before finishing the discussion of these case studies. Figure C3.1.57 shows a pair of optical probes used for on-line measurements of liquid layer thickness at a coating station. Apart from operating in liquid films, the method is essentially similar in nature to the ones described above, so no more technical details will be given.

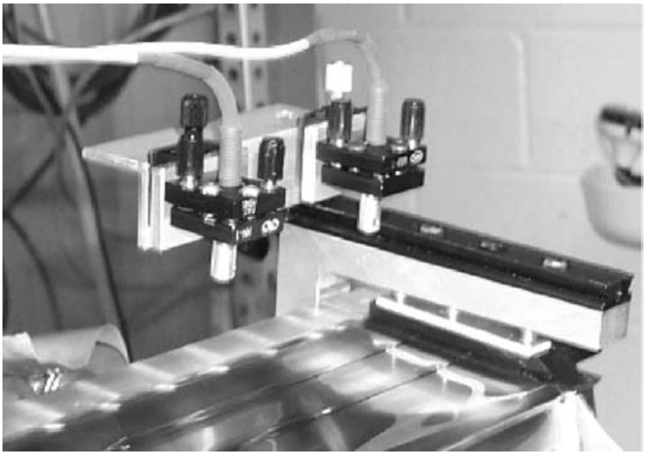


Figure C3.1.57. Photograph showing dual probes for on-line assessment of liquid layer thickness during operation of a coating plant.

These probes are invaluable aids to assess coating uniformity, and coating dynamics. They have been utilized to assess the time it takes a production process to settle down and reach steady state after the production fluids are first introduced into the coating stations.

C3.1.7 Conclusions

This chapter has given an overview of the field of fibre sensors, starting with simple intensity types, then more complex interferometric ones. Designs for multiplexed and fully distributed sensors have been dealt with in considerable depth. The chapter has concluded with an extensive case study to illustrate how a particular type of sensor can be used and developed for use in industrial process monitoring. Owing to the breadth of the subject, even this large chapter could not cover many concepts, so the authors apologize to any researchers if their work has been omitted. A comprehensive bibliography is given below in order to allow the reader to glean further details of this subject.

It has been shown that there are very many ways of sensing using optical fibres. Many sensors reported in scientific papers will retain only academic interest, but a few technologies are already making significant inroads into industrial, medical and military applications. The general trend, as with any relatively new technology, is first to fill niche markets, before costs can eventually be reduced to meet mass markets. Although more complex in nature, the use of techniques for achieving multiplexed and distributed sensors is increasing rapidly. It is difficult to generalize on preferred approaches, but the TDM and WDM methods of multiplexing, and OTDR-type methods (including Raman, Brillouin and fluorescent variants) for distributed sensing have many advantages. The achievement of more sensors per monitoring station, and the correspondingly greater ease of making comparisons of the measurand value at each sensor head, is a factor likely to increase the practical use of optical sensor systems. The major area requiring evolution of technology is the extension of these methods to address (and separate effectively!) several different physical parameters simultaneously on a common optical highway in an economic manner.

It will be interesting to observe which methods endure the passage of time to be developed into cost-effective and reliable system instrumentation. The factors in favour of optical sensor highways are the ever-decreasing cost of fibre cable and the dramatic improvements of cost and performance of optical components.

References

[1] Brabant J M, Moyer B J and Wallace R 1957 Lead glass Cerenkov radiation photon spectrometer *Rev. Sci. Instrum.* **28** 421

[2] England J B A 1976 Detection of ionizing radiations *J. Phys. E (Sci. Instrum.)* **9** 233–251

[3] Fenyves E J (conference chair) 1994 Scintillating fibre technology and applications II *Proc. SPIE* vol 2281 (several relevant papers in these proceedings)

[4] Fields J N 1979 Attenuation of a parabolic index fibre with periodic bends *Appl. Phys. Lett.* **36** 779–801

[5] Fields J N and Cole J H 1980 Fibre microbend acoustic sensor *Appl. Opt.* **9** 3265–3267

[6] Berthold J W and Reed S E 1979 Microbend fibre-optic strain gauge *US Patent No* 5020379

[7] Harmer A L 1981 *UK Patent* 1584173

[8] Harmer A L 1986 *US Patent* 4618764

[9] Michie W C, Culshaw B, McKenzie I, Moran C, Graham N B, Santos F, Gardiner P T, Berqqvist E and Carlstrom B 1994 A fibre optic hydrogel probe for distributed chemical measurements, Proc. 10th Optical Fibre Sensors Int. Conf., Glasgow *Proc. SPIE* **2360** pp130–133

[10] Michie W C, Culshaw B, McKenziei I, Konstantakis M, Graham N B, Moran C, Santos F, Bergqvist B and Carlstrom B 1995 Distributed sensor for water and pH measurement using fibre optic and swellable polymeric materials *Opt. Lett.* **20** 103–105

[11] Gaebler W and Braunig D 1983 Application of optical fibre waveguides in radiation dosimetry *Proc. First Int. Conf. on Optical Fibre Sensors, London* pp 185–189

[12] Pinchbeck D and Kitchen C 1985 Optical fibres for cryogenic leak detection *Proc. Conf. on Electronics in Oil and Gas, London* (London: Cahners Exhibits Ltd) pp 469–501

[13] Harmer A L 1982 Optical fibre sensors and instrumentation—a multi-client study *Interim Report, by Harmer, for Battelle Geneva Research Laboratory*

[14] Dakin J P and Kahn D A 1977 A novel fibre optic temperature probe *Opt. Quantum Electron.* **9** 540–544

[15] Diles R R 1983 High temperature, optical fibre thermometer *J. Appl. Phys.* **54** 1198–1201

[16] Pitt G D 1982 *Electron. Commun.* **57** 102–105

[17] Dakin J P and Holliday M G 1983 A liquid level sensor, based on O–H or C–H absorption monitoring *Proc. First International Optical Fibre Sensors Conf., OFS1, London*

[18] Frank W E 1966 Detection and measurement device having a small flexible fibre transmission line *US Patent* 3273447

[19] Menadier C, Kissinger C and Adkins H 1967 The fotonic sensor *Instrum. Contr. Syst.* **40** 114–120

[20] Culshaw B, Foley J and Giles I P 1984 A balancing technique for optical fibre intensity modulated transducers, Proc. Second International Optical Fibre Sensors conf, OFS2, Stuttgart *Proc. SPIE* **574** 117–120

[21] Culshaw B 1987 Optically excited resonant sensors *Institute of Physics (UK) Short Meeting Series No 7* pp 33–44

[22] Culshaw B 1989 Silicon in optical fibre sensors *Optical Fibre Sensors* vol 2, eds J P Dakin and B Culshaw ISBN 0-89006-376-1 (Boston, MA: Artech House) chapter 13, pp 475–509

[23] Meltz G, Morey W W and Glenn W H 1989 Formation of Bragg gratings in optical fibres by a transverse holographic method *Opt. Lett.* **14** 823–825

[24] Morey W W, Meltz G and Glenn W H 1989 Fibre optic Bragg grating sensors, Proc. Fibre Optic and Laser Sensors VII *Proc. SPIE* **1169** 98–106

[25] St J Russell P and Archambault J L 1996 Fibre gratings *Optical Fibre Sensors* vol 3, eds J P Dakin and B Culshaw ISBN 0-89006-932-8 (Boston, MA: Artech House) chapter 2, pp 475–509

[26] Morey W W, Dunphy J R and Meltz G 1991 Multiplexing fibre Bragg grating sensors, Proc. First Int. Conf. on Distributed and Multiplexed Fibre Optic Sensors, Boston, ed A D Kersey and J P Dakin *Proc SPIE* **1586** 216–224

[27] Kersey A D 1993 Interrogation and multiplexing techniques for fibre Bragg grating strain sensors, Proc. Distributed and multiplexed fibre optic sensors III, Boston, USA, 1993 *SPIE* **2071** 30–48

[28] Xu M G, Archambault J L, Reekie L and Dakin J P 1994 Structural bending gauge using fibre gratings *Proc. Fibre Optic and Laser Sensors XII, San Diego, SPIE* vol 2292 (Paper 48)

[29] Measures R M, Melle S M and Liu K 1992 Wavelength demodulated Bragg grating fibre optic sensing systems for addressing smart structure critical issues *Smart Mater. Struct.* **1** 36–44

[30] Davies M A, Bellemore D G and Kersey A D 1994 Structural strain mapping using a wavelength/time division addressed fibre Bragg grating array *Proc. Second European Conf. on Smart Structures and Materials, Glasgow, SPIE* vol 2361 pp 342–345

[31] Nellen P M, Broennimann R, Sennhauser U J, Askins C G and Putman M A 1995 Applications of distributed fibre Bragg grating sensors in civil engineering, Proc. Distributed and Multiplexed Fibre Optic Sensors V, Munich, ed A D Kersey and J P Dakin *Proc SPIE* **2507** 14–24

[32] Jackson D A, Ribeiro A B L, Reekie L, Archambault J L, St J Russell P 1993 Simultaneous interrogation of fibre optic grating sensors *Proc. Ninth Int. Optical Fibre Sensors Conference, OFS9, Florence* pp 39–42

[33] Kersey A D, Berkoff T A and Morey W W 1993 Fibre Fabry–Perot demodulation for Bragg grating strain sensors *Proc. Ninth International Optical Fibre Sensors Conference, OFS9, Florence* pp 39–42

[34] Koo K P and Kersey A D 1995 Fibre laser sensor with ultra-high strain resolution, using interferometric interrogation *Electron. Lett.* **31** 1180–1182

[35] Xu M G, Geiger H and Dakin J P 1996 Modelling and performance analysis of a fibre Bragg grating interrogation system using an acousto-optic tunable filter *J. Lightwave Technol.* **14** 391–396

[36] Ecke W, Schauer J, Usbeck K, Willsch R and Dakin J P 1997 Improvement of the stability of fibre grating interrogation systems, using active and passive polarisation scrambling devices *Proc. 12th Int. Optical Fibre Sensors Conf., OFS12, Williamsburg, USA* pp 484–487

[37] Hadeler O, Richardson D J and Dakin J P 1999 DFB fibre laser structure for simultaneous strain and temperature measurements in concrete structures, Proc. of Int. Conf. Smart Systems for Bridges Structures and Highways *SPIE* **3670** 332–341

[38] Crickmore R I, Gunning M J, Stefanov J and Dakin J P 2003 Beat frequency measurement system for multiple dual polarisation fibre DFB lasers *IEEE Sensors J.* **3** 115–120

[39] Kyuma K, Tai S, Sawada T and Nunoshita M 1982 Fibre optic instrument for temperature measurement *IEEE J. Quantum Electron.* **QE-18** 676–679 (see also chapter 17, by Kyuma, in vol 2 of bibliography 1)

[40] Lee C E and Taylor H F 1991 Fibre optic Fabry–Perot temperature sensor using a low-coherence light source *J. Lightwave Technol.* **LT-9** 129–134

[41] Ovren C, Adolfson M and Hok B 1983 Fibre optic systems for temperature and vibration measurements in industrial applications *Proc. Optical Techniques in Process Control Den Hague* (Cranfield University and BHRA Fluid Engineering) pp 67–81 (See also ASEA Innovation data sheet for semiconductor temperature measurement, 1986)

[42] McCormack J S 1981 Remote optical measurement of temperature, using luminescent materials *Electron. Lett.* **17** 630–632

[43] Sholes R R and Small J G 1980 Fluorescent decay thermometer, with biological applications *Rev. Sci. Instrum.* **51** 882

[44] Gratton K T V and Palmer A W 1986 Fluorescent monitoring for optical temperature sensing *Fibre Opt., SPIE* **630** 256–265

[45] Peterson J I, Goldstein S R, Fitzgerald R V and Buckhold D K 1980 Fibre optic pH probe for physiological use *Anal. Chem.* **52** 864

[46] Lippitsch M E, Pusterhofer J, Leiner M J P and Wolfbeiss O S 1988 Fibre optic oxygen sensor with the decay time as the information carrier *Anal. Chim. Acta* **205** 1

[47] Jackson D A and Jones J D C 1989 Interferometers *Optical Fibre Sensors* vol II, eds J P Dakin and B Culshaw (Boston, MA: Artech House) chapter 10, pp 329–380

[48] Hotate K 1997 Fibre optic gyros *Optical Fibre Sensors* vol IV, eds J P Dakin and B Culshaw (Boston, MA: Artech House) chapter 11, pp 167–206

[49] Hotate K 2000 Fibre optic gyros *Trends in Optical Nondestructive Testing and Inspection* (Amsterdam: Elsevier Science) chapter 32, pp 487–502

[50] Lefèvre H C 1989 Fibre optic gyroscope *Optical Fibre Sensors* vol II, eds J P Dakin and B Culshaw (Boston, MA: Artech House) chapter 11, pp 381–429

[51] Sagnac G 1914 Effet tourbillonnaire optique. La circulation de l'ether lumineux dans un interferographe tournant *J. Phys. Radium* **4** 177–195

[52] Willsch M and Bosselmann T 2002 Optical current sensor application in the harsh environment of a 120 MVA power generator *Proc. 15th Int. Conf. Optical Fibre Sensors, OFS 2002, Portland, Paper ThD2* pp 407–410

[53] Nelson A R, McMahon D H and Gravel R L 1980 Passive multiplexing system for fibre optic sensors *Appl. Opt.* **19** 2917–2920

[54] Thayer D R, Lyons P B, Looney L D, Manning J P and Malone R M 1981 Preparation, installation and calibration of a 152 fibre imaging experiment at the Nevada test site *SPIE* **296** 191–194

[55] Dakin J P and King A J 1983 Limitations of a single optical fibre fluorimeter system due to background fluorescence *Proc. First Int. Conf. on Optical Fibre Sensors, London (IEE Conf. Pub. 221)* pp 195–199

[56] Miller G E and Lindsay T A 1978 Feasibility demonstration of fibre optic digital status monitoring devices *Boeing Aerospace Company (Seattle) Report No* D296-10048-1

[57] James K A, Quick W H and Strachan V H 1979 Fibre optics: the way to true digital sensors *Contr. Eng.* 30–33

[58] Nelson A R, McMahon D H and Van de Vaart H 1981 Multiplexing system for fibre optic sensors using pulse compression techniques *Electron. Lett.* **17** 263–264

[59] Barnoski M K and Jensen S M 1976 Fibre waveguides: a novel technique for investigating attenuation characteristics *Appl. Opt.* **15** 2112–2115

[60] Personick S D 1977 Photon probe-an optical fibre time domain reflectometer *Bell Systems Tech. J.* **56** 355–366

[61] Desforges F X, Graindorge P, Jeunhomme L B and Arditty H J 1986 Progress in OTDR optical fibre sensor networks *SPIE 718, Paper 31*

[62] Asawa C K, Yao S K, Stearns R C, Mota N L and Downs J W 1982 High-sensitivity fibre-optic strain sensors for measuring structural distortion *Electron. Lett.* **18** 362–364

[63] Bruinsma A J A, Van Zuylen P, Lamberts C W and de Krijger A J T 1984 Fibre optic strain measurement for structural integrity monitoring *Proc. Second Int. Conf. Optical Fibre Sensors, OFS'84, Stuttgart* (Berlin: VDE) pp 399–401

[64] Claus R O, Jackson B S and Bennett K D 1985 Nondestructive testing of composite materials by OTDR in imbedded optical fibres *Proc. Fibre Laser Sensors III. SPIE* **566** 243–248

[65] Skolnik M I 1970 *Radar Handbook* (New York: McGraw-Hill)

[66] Dakin J P, Wade C A and Henning M 1984 Novel optical fibre hydrophone array using a single laser source and detector *Electron. Lett.* **20** 51–53

[67] Wade C A 1982 *British Patent Application No* 8207961 (Priority date 18 March 1982)

[68] Dakin J P and Wade C A 1982 *British Patent Application No* 8220793

[69] Dakin J P, Wade C A and Ellis G 1986 A novel 3-wave mixing approach to coherent communications *Proc. 12th ECOC, Barcelona (post deadline paper)*

[70] Nash P, Latchem J, Cranch G, Motley S, Bautista A, Kirkendall K, Dandridge A, Henshaw M and Churchill J 2002 Design, development and construction of fibre-optic bottom-mounted array *Proc OFS 15, 15th Int. Conf. on Optical Fiber Sensors, Portland, Oregon, USA, IEEE 02Ex533*

[71] Cranch G, Kirkendall K, Daley K, Motley S, Bautista A, Salerno J, Nash P, Latchem J and Crickmore R 2003 Large scale remotely pumped and interrogated fiber-optic interferometric sensor array *IEEE Photon. Technol. Lett.* May 2003

[72] Pendleton-Hughes S, Weston N and Carter A C 1985 Forty channel wavelength multiplexing for short-haul wideband communications networks *Proc. 11th European Conf on Optical Communications (ECOC 85), Venice* pp 649–652

[73] Russell S J and Dakin J P 1999 Location of time-varying strain disturbances over a 40km fibre section, using a dual-Sagnac interferometer with a single source and detector, Proc. OFS-13, Kyongju, Korea *SPIE* **3746** 580–583

[74] Dakin J P and Liddicoat T J 1982 A wavelength multiplexed optical shaft encoder *Meas. Contr.* **15** 176–177

[75] Lear R D 1981 Time dependent recordings of images transmitted over optical fibres, Proc. Conf. on Fibre Optics in Adverse Environments, San Diego *SPIE* **296** 228–233

[76] Vohra S, Dandridge A, Danver B and Tveten A 1996 A hybrid WDM/TDM reflectometric array *Proc. 11th Int. Conf. on Optical Fibre Sensors, OFS96, Sapporo, Japan* pp 534–537

[77] Davies M A and Kersey A D 1995 Application of a fibre Fourier transform spectrometer to the detection of wavelength-encoded signals from Bragg grating sensors *IEEE J. Lightwave Technol.* **13** 1289–1295

[78] Dakin J P and Volanthen M 1999 Review of distributed and multiplexed fibre grating sensors and discussion of problem areas (Invited) *Proc. Photonics East Int. Conf., Boston, 1999, Proc SPIE* vol 3860 (paper 16)

[79] Mlodzianowski J, Uttam D and Culshaw B 1986 A multiplexed system for analogue point sensors *Proc. IEE Colloq. on Distributed Optical Fibre Sensors (IEE Digest No 86/74)* (London: IEE) pp 12/1–12/3

[80] Uttam D and Culshaw B 1982 Optical FM applied to coherent interferometric sensors *IEE Colloq. on Optical Fibre Sensors (IEE Digest No 1982/60)* (London: IEE)

[81] Giles I D, Uttam D, Culshaw B and Davies D E N 1983 Coherent optical fibre sensors with modulated laser sources *Electron. Lett.* **19** 14–15

[82] Al Chalabi S A, Culshaw B, Davies D E N, Giles I P and Uttam D 1985 Multiplexed optical fibre interferometers: an analysis based on radar systems *Proc. IEE* **132** 150–156

[83] Sakai I 1986 Frequency-division multiplexing of optical fibre sensors using a frequency-modulated source *Opt. Quantum Electron.* **18** 279–289

[84] Mallalieu K I, Youngquist R and Davies D E N 1986 RF-band FMCW for passive multiplexing of multimode fibre optic sensors *Proc. IEE. Colloq. on Distributed Optical Fibre Sensors (IEE Digest No 1986/74)* (London: IEE) pp 4/1–4/3

[85] Brooks J L, Wentworth R H, Youngquist R C, Tur M, Kim B Y and Shaw H J 1983 Coherence multiplexing of fibre-optic interferometric sensors *J. Lightwave Technol.* **LT3** 1062–1072

[86] Dakin J P, Ecke W, Rothhardt M, Schauer J, Usbeck K and Willsch R 1997 New multiplexing scheme for monitoring fibre optic Bragg grating sensors in the coherence domain (Invited) *Proc. OFS 12 Int. Conf., Williamsburg, USA* pp 31–34

[87] Kingsley S A, Davies D E N, Culshaw B and Howard D 1978 Fibredyne systems *Proc. Fibre Optic Communications Conf., Chicago* (Information Gatekeepers Inc.)

[88] Culshaw B, Ball P R, Pond J C and Sadler A A 1981 Optical fibre data collection *Electron. Power* 148–150

[89] Crossley S D, Giles I P, Culshaw B and Sadler A A 1983 An optical fibre data telemetry system for use in remote or hazardous locations. Proc. Int. Conf. on Optical Techniques in Process Control, The Hague (Cranfield: BHRA) pp 121–132 *Appl. Opt.* **15** 2112–2115

[90] Dakin J P and Pratt D J 1986 Fibre-optic distributed temperature measurement—a comparative study of techniques *Proc. IEE Colloq. on Distributed Optical Fibre Sensors (IEE Digest No 1986/74)* (London: IEE) pp 10/1–10/4

[91] Rogers A J 1980 Polarisation optical time domain reflectometry *Electron. Lett.* **16** 489–490

[92] Di Vita P and Rossi U 1980 The backscattering technique: its field of applicability in fibre diagnostics and attenuation measurements *Opt. Quantum Electron.* **11** 17–22

[93] Conduit A J, Payne D N, Hartog A H and Gold M P 1981 Optical fibre diameter variations and their effect on backscatter loss measurement *Electron. Lett.* **17** 308–310

[94] Theocharous E 1983 Differential absorption distributed thermometer *Proc. First Int. Conf. on Optical Fibre Sensors* (London: IEE) pp 10–12

[95] Quoi K W, Lieberman R, Cohen L G, Shenk D S and Simpson J R 1992 Rare-earth-doped optical fibres for temperature sensing *IEEE J. Lightwave Technol.* **10** 847–851

[96] Hardy E E, David D J, Kapany N S and Unterleitner F C 1975 Coated optical guides for spectrophotometry of chemical reactions *Nature* **257** 666

[97] Blyler L L, Ferrara J A and MacChesney J B 1988 A plastic clad silica fibre chemical sensor for ammonia *Proc OFS88 Technical Digest* Opt. Soc. Am.) p 369–373

[98] Hartog A H, Payne D N and Conduit A J 1980 POTDR: experimental results and application to loss and birefringement measurements in single mode fibres *Proc. Sixth European Conf. on Optical Communication, ECOC80, York* (post deadline paper) (London: IEE)

[99] Kim B Y and Choi S S 1981 Backscattering measurements of bending-induced birefringence in single mode fibres *Electron. Lett.* **17** 193–195

[100] Ross J N 1981 Measurement of magnetic field by POTDR *Electron. Lett.* **11** 596–597

[101] Rogers A J 1981 POTDR a technique for the measurement of field distributions *Appl. Opt.* **20** 1060–1074

[102] Healey P 1984 Fading in heterodyne OTDR *Electron. Lett.* **20** 30–32

[103] Hartog A H and Payne D N 1982 Fibre optic temperature distribution sensor *Proc. IEE Colloq. Optical Fibre Sensors* (London: IEE)

[104] Dakin J P, Pratt D J, Bibby G W and Ross J N 1985 Distributed anti-Stokes ratio thermometry *Proc. Third Int. Conf: on Optical Fibre Sensors, San Diego (post-deadline paper)*

[105] Dakin J P, Pratt D J, Bibby G W and Ross J N 1985 Distributed optical fibre Raman temperature sensor using a semiconductor light source and detector *Electron. Lett.* **21** 569–570

[106] Dakin J P 1984 *UK Patent Application GB* 2156513A (published 9 October 1985)

[107] Poole S B, Payne D N and Fermann M E 1985 Fabrication of low-loss optical fibres containing rare-earth ions *Electron. Lett.* **21** 737–738

[108] Lieberman R A, Blyler L L and Cohen L G 1990 A distributed fibre optic sensor based on cladding fluorescence *IEEE J. Lightwave Technol.* **8** 212–220

[109] Horiguchi T and Tateda M 1989 Optical fibre attenuation investigation using stimulated Brillouin scattering between a pulse and a continuous wave *Opt. Lett.* **14** 408–410

[110] Horiguchi T and Tateda M 1989 BOTDA—non-destructive measurement of single mode optical fibre attenuation investigation characteristics using Brillouin interaction: theory *J. Lightwave Technol.* **7** 1170–1176

[111] Horiguchi T, Kurashima T and Tateda M 1990 A technique to measure distributed strain in optical fibres *IEEE Photon. Technol. Lett.* **2** 352–354

[112] Horiguchi T, Shimizu K, Kurashima T and Koyamada Y 1995 Advances in distributed sensing techniques using Brillouin scattering, Proc. Distributed and Multiplexed Fibre Optic Sensors II, Boston, USA, ed J P Dakin and A D Kersey *SPIE* **2507** 126–135

[113] Wait P C and Newson T P 1996 Landau–Placzek ratio, applied to distributed sensing *Opt. Commun.* **122** 141–146

[114] Maughan S L, Kee H H and Newson T P N 2001 Simultaneous distributed fibre temperature and strain sensor, using microwave coherent detection of spontaneous Brillouin backscatter *Meas. Sci. Technol.* **12** 834–842

[115] Eickhoff W and Ulrich R 1981 Optical frequency domain reflectometry in single mode fibre *Appl. Phys. Lett.* **39** 693–695

[116] Kingsley S A and Davies D E N 1985 OFDR diagnostics for fibre and integrated-optic systems *Electron. Lett.* **21** 434–435

[117] Ghafoori-Shiraz H and Okoshi T 1986 Fault location in optical fibres using optical-frequency-domain reflectometry *J. Lightwave Technol.* **LT-4** 316–322

[118] Hotate K 1998 Fibre sensor technology today (Invited) *Optical Fibre Technology* (London: Academic Press) pp 356–402

[119] Hotate K 1999 Coherent photonic sensing *Sensors Update* vol 6 (Wiley-VCH) chapter 8, pp 131–162

[120] Hotate K and Ong S S L 2003 Distributed dynamic strain measurement using a correlation-based Brillouin sensing system *IEEE Photon. Technol. Lett.* **12** 272–274

[121] Hotate K 2002 Application of synthesized coherence function to distributed optical sensing *IOP Meas. Sci. Technol.* **13** 1746–1755

[122] Franks R B, Torruellas W and Youngquist R C 1985 Birefringent stress location sensor *SPIE*, **586** 84–88

[123] Dakin J P, Pearce D A, Strong A P and Wade C A 1987 A novel distributed optical fibre sensing system enabling location of disturbances in a Sagnac loop interferometer *Proc. SPIE* **838** paper 18

[124] Dakin J P, Pearce D A, Strong A P and Wade C A 1988 A novel distributed optical fibre sensing system enabling location of disturbances in a Sagnac loop interferometer *Proceedings EFOC/LAN 88* Information Gatekeepers Inc.) pp 276–279

[125] Udd E 1991 Sagnac distributed sensor concepts *Proc. SPIE* **1586** 46–52

[126] Spammer S J, Swart P L and Boosen A 1996 Interferometric distributed fibre optical sensor *Appl. Opt.* **35** 4522–4523

[127] Spammer S J, Chtcherbakov A A and Swart P L 1996 Dual wavelength Sagnac–Michelson distributed optical fibre sensor *Proc. SPIE* 2834–2838

[128] Spammer S J, Chtcherbakov A A and Swart P L 1998 Distributed dual wavelength Sagnac impact sensor *Microwave Opt. Technol. Lett.* **17** 170–173

[129] Ronnekleiv E, Blotekjaer E K and Kranes K 1993 Distributed fibre sensor for location of disturbances *Proceedings OFS-9 Int. Conf., Paper PD7*

[130] Fang X J 1996 Fibre-optic distributed sensing by a two-loop Sagnac interferometer *Opt. Lett.* **21** 444–446

[131] Dandridge A, Tveten A B and Giallorenzi T G 1982 Homodyne demodulation scheme for fibre optic sensors using phase generated carrier *IEEE Trans. Microwave Theory Tech.* **MTT-30** 1635–1641

[132] Farries M C and Rogers A J 1984 Distributed sensing using stimulated Raman interaction in a monomode optical fibre *Proc. Second Int. Conf: Optical Fibre Sensors, OFS'84, Stuttgart* (Berlin: VDE) pp 121–132

[133] Martin J and Ouellette F 1994 Novel writing technique of long and highly-reflective in-fibre gratings *Electron. Lett.* **30** 811–812

[134] Huang S H, Ohn M M, Leblanc M, Lee R and Measures R M 1994 Fibre optic intra-grating distributed strain sensor, Proc Int. Conf. Distributed and Multiplexed Fibre Optic Sensors IV, San Diego, ed A D Kersey and J P Dakin *Proc SPIE 2294* 81–92

[135] Labelet P, Fonjallaz P Y, Limberger H G, Salathe R P, Zimmer C and Gilgen H H 1993 Bragg grating characterisation by optical low-coherence reflectometry *IEEE Photon. Technol. Lett.* **5** 565–567

[136] Volanthen M, Geiger H, Cole M J, Laming R I and Dakin J P 1996 Low coherence technique to characterise reflectivity and time delay, as a function of wavelength, within a long Bragg grating *Electron. Lett.* **32** 757–758

[137] Volanthen M, Geiger H, Cole M J and Dakin J P 1996 Measurement of arbitrary strain profiles within fibre gratings *Electron. Lett.* **32** 1028–1029

[138] Volanthen M, Geiger H and Dakin J P 1997 Low-coherence grating characterisation scheme *Proc. IEE Colloquium on Optical Fibre Gratings, London*

[139] Volanthen M, Geiger H and Dakin J P 1997 Distributed grating sensors using low coherence reflectometry *IEEE J. Lightwave Technol.* **15** 2076–2082

[140] Bush J, Davis P and Marcus M A 2001 All-fibre coherence domain interferometric techniques *Proc Int. Conf. Fibre Optic Sensor Technology II, Proc. SPIE* **4204** 71–80

[141] Danielson B and Boisrobert C 1991 Absolute optical ranging using low coherence interferometry *Appl. Opt.* **30** 2975

[142] Harris H W, Lee J-R and Marcus M A 1999 Noncoherent light interferometry as a thickness gauge, ed M A Marcus and A Wang *Proc. SPIE* **3538**

[143] Lee J-R and Marcus M A 2000 Method for determining the retardation of a material using non-coherent light interferometry *US Patent No* 6,034,774

[144] Marcus M A, Gross S and Wideman D 1997 Associated dual interferometric measurement apparatus for determining a physical property of an object *US Patent No* 5,659,392

[145] Marcus M A, Trembley T and Uerz D 1998 Digital camera image sensor positioning apparatus including a non-coherent light interferometer *US Patent No* 5,757,486

[146] Marcus M A, Dilella E A, Lee J-R, Lowry D R and Trembley T M 2003 Measurement method and apparatus of an external digital camera imager assembly *US Patent* 6,512,587 B1

[147] Bracewell R 1978 *The Fourier Transform and Its Applications* 2nd edn (New York: McGraw-Hill)

[148] Bruinsma A J A and Jongeling T J M 1989 Some other applications for fibre optic sensors *Optical Fibre Sensors* vol II, eds J P Dakin and B Culshaw (Boston, MA: Artech House) chapter 19, pp 721–765

[149] Dakin J P and Pratt D J 1985 Improved non-invasive fibredyne highway based on polarimetric detection of strain in polarisation-maintaining fibres *Electron. Lett.* **21** 1224–1225

[150] Davies M A and Kersey A D 1994 All-fibre Bragg grating strain sensor demodulation technique, using a wavelength-division coupler *Electron. Lett.* **30** 75–77

[151] Kurashima T, Horiguchi T and Koyamada Y 1992 Measurement of temperature and strain distribution by Brillouin shift in silica optical fibres, Proc. Distributed and Multiplexed Fibre Optic Sensors II, part of O/E Fibres'92, Boston, USA, ed J P Dakin and A D Kersey *SPIE* **1797** 2–13

[152] Jones B E 1981 Simple optical sensors for the process industries, using incoherent light *Proc. Inst. of Measurement and Control Symp. on Optical Sensors and Optical Techniques in Instrumentation, London* (London: IMC)

[153] Kersey A 1997 Multiplexing techniques for fibre-optic sensors *Optical Fibre Sensors* vol IV, eds J P Dakin and B Culshaw (Boston, MA: Artech House) chapter 15, pp 369–407

[154] Rogers A J 1985 Intrinsic and extrinsic distributed optical fibre sensors *SPIE* **586** 51–52; *SPIE* **566** 234–242

[155] Kurashima T, Horiguchi H, Izumita H, Furukawa S and Koyamada Y 1993 Brillouin optical-fiber time domain reflectometry *IEICE Trans. Commun.* **E76-B** 382

Bibliography

1. Dakin J P and Culshaw B (eds) 1988, 1989, 1996, 1997 (in order of volume number), *Optical Fibre Sensors* vols 1–4 (Boston: Artech House), ISBN 0-89006-317-6, 0-89006-376-1, 0-89006-932-8, 0-89006-940-9
2. Wolfbeiss O O *Fibre Optic Chemical Sensors and Biosensors* vols 1 and 2 (Boston: CRC Press), ISBN 08493-5508-7, 08493-5509-5
3. Udd E (ed) 1995 *Fibre Optic Smart Structures* (New York: Wiley), ISBN 0-471-5548-0
4. Udd E Organiser of SPIE series of video tutorial courses on optical fibre sensors
5. Dakin J P and Culshaw B 1986 Distributed fibre optic sensors *SPIE Crit. Rev.* **CR 44**
6. Davies D E N 1984 Signal processing for distributed optical fibre sensors *Proc. Int. Conf on Optical Fibre Sensors, OFS84, Stuttgart* pp 285–295
7. Kersey A 1997 Multiplexing techniques for fibre-optic sensors *Optical Fibre Sensors* vol IV, ed J P Dakin and B Culshaw (Boston: Artech House) chapter 15, pp 369–407
8. Kersey A D 1993 Interrogation and multiplexing techniques for fibre Bragg grating strain sensors, Proc. Distributed and Multiplexed Fibre Optic Sensors III, Boston, USA, 1993 *SPIE* **2071** 30–48
9. Kersey A, Davies M A, Patrick H J, LeBlanc M, Koo K P, Askins C G, Putman M A and Friebele E J 1997 Fibre grating sensors *J. Lightwave Technol.* **15** 1442–1462
10. Rogers A J 1985 Intrinsic and extrinsic distributed optical fibre sensors *SPIE* **586** 51–52; *SPIE* **566** 234–242
11. Rogers A J 1988 Distributed optical fibre sensors for the measurement of pressure strain and temperature *Phys. Rep.* **169** 99–143
12. Culshaw B 1986 Distributed and multiplexed fibre optic sensor systems *Proc. 11th Course of the Int. School of Quantum Electronics, Erice, Sicily* (Dordrecht: Martinus Nijhoff)

C3.2

Remote optical sensing by laser

J. Michael Vaughan

C3.2.1 Introduction—sensing by laser radar (LIDAR)

C3.2.1.1 Laser radar and laser properties

The myriad of modern applications for lasers includes the remote sensing and investigation of distant objects. When this work is conducted outdoors, at ranges of a few tens of metres to many hundreds of kilometres, the subject is usually called 'laser radar'. It must be admitted that this is something of a misnomer—the term radar itself derives from 'radio detection and ranging', which was of course developed over 60 years ago for detection of aircraft by long-wavelength radio waves. However, we are presently considering the use of very much shorter wavelengths—that is, light waves in the visible and near-visible region. In consequence, the terms 'lidar' for 'light detection and ranging' and 'ladar' for 'laser detection and ranging' have also been introduced. While some attempts have been made to differentiate the usage of the three expressions 'laser radar', 'lidar' and 'ladar', they are in fact generally used freely and interchangeably. However, one can do a great deal more than just 'detect and range' with lasers. In consequence, two other terms are also used for more specific applications: laser Doppler velocimetry (LDV)—use of the Doppler principle for remote velocity measurements, and differential absorption by laser (DIAL)—chemical detection with lasers tuned on and off resonance absorption.

The principle of laser radar is of course very simple and straightforward. The laser beam is sent out from an optical transmitter towards the object of interest, often referred to as the 'target'. The target will scatter light from the beam and some of this will be reflected back towards a receiver—usually placed adjacent to the transmitter and often indeed using the same optical arraFngement set up as a transmitter/receiver. This scattered and reflected light will contain information about the target so that, after the light has been detected within the receiver and converted into an electrical signal, the information may be extracted. Lasers have many very desirable properties for use in remote sensing and these are summarized in table C3.2.1. Many of these laser properties and the unique characteristics of laser radiation are discussed at considerable length in other chapters of this volume, but it is worth noting a few points of particular relevance to remote sensing.

Many of the secondary characteristics derive from the primary characteristics; thus, precise focusing and pointing accuracy are due to good coherence properties and, combined with high brightness, can contribute enormous power density (it has been noted that, in certain high energy, exceptionally short-pulse lasers, the actual instantaneous power level may be greater than all the power stations of the world combined). Of particular value for lidar operation is the wide range of available lasers and wavelengths, extending from the near ultra-violet at wavelength $\lambda \sim 0.3\,\mu\mathrm{m}$ up to the mid infra-red at $\lambda \sim 12\,\mu\mathrm{m}$. For outdoors, any laser system must of course be operated in a way that presents no hazard whatsoever to people or equipment in the locality. At wavelengths between ~ 0.38 and

Table C3.2.1. Laser properties.

Primary characteristics	High brightness Good spatial coherence Good temporal coherence
Secondary characteristics	Wide range of available wavelength Continuously tuneable in certain regions Continuous wave (cw) and pulsed output Great pointing accuracy Precise focusing Enormous power density

$\sim 1.5\ \mu m$, this necessitates careful control of the transmitted beam to ensure laser intensity levels remain below established thresholds (taking account of scintillation in the atmosphere; see, e.g. Ref. [1]). In the so-called eye-safe regime ($\lambda < 0.38$ and $> 1.5\ \mu m$), where radiation is at least not transmitted through to the retina, conditions are not quite so stringent. Thus, for many applications, operation with such eye-safe lasers has been preferred. This in fact has rather profound implications for the development of lidar technology and its broad division between the two classes of laser radar discussed in the following sections.

C3.2.1.2 Remote sensing applications of laser radar

A list of actual and potential remote sensing applications for lasers is given in table C3.2.2. For such a great range of tasks, a wide variety of principles, techniques and systems have been applied and developed; some of the more important aspects of these are outlined in the following two sections. Following such basic considerations, there are a number of possible ways of discussing their application to remote-sensing problems; one could, for example, provide a categorization based on the lidar base from fixed or mobile ground platforms, from the air (by balloon or aircraft) or from a space craft. However, a categorization based on the field of application seems likely to be more informative. Accordingly, in the present report, lidar remote sensing is discussed under six separate headings: atmospheric sensing; problems in aviation; chemical and pollution studies; military applications; geoscience; and measurements from space. Obviously, most of these individual topics overlap across several of the items listed in table C3.2.2.

The present chapter deals with active remote sensing by laser. For an account of passive sensing with thermal emission in the infrared, see, for example, Chapter B9 in this volume. For a short list of

Table C3.2.2. Some remote-sensing applications of laser radar.

Wind measurement—meteorology, airfield and aircraft shear and turbulence warning, aircraft wake vortices
Atmospheric measurement—clouds, precipitation, aerosols (dust), temperature, pressure, dispersion
Chemical species, pollutant and gas detection
Airborne operation—avionics data, true airspeed, obstacle and terrain avoidance
Scene and object imaging with various discriminants—intensity, glint, polarization, frequency shift, etc
Range and three-dimensional sizing, depth sounding
Vibrational analysis of surfaces and structures
Spaceborne measurements—cloud, atmospheric scattering, global wind field

further reading including books, reviews and compendia of papers, see the Reference section at the end of this chapter.

C3.2.2 Some fundamental considerations in laser radar

C3.2.2.1 Basic principles and phenomena for remote sensing with lasers

Some of the basic principles and phenomena that may be utilized in laser radar are indicated in table C3.2.3. The scattering of light from the target is of course fundamental to the technique and has many possible forms. Of those noted, elastic (without gross change of frequency), specular (mirror-like) and diffuse (as from a rough surface) are terms relating to surface scattering. Mie refers to scattering from particles of size comparable with the wavelength λ. Rayleigh and Brillouin scattering arise from density fluctuations due to thermal effects at the surface and (more strongly) within the bulk of a material, whereas Raman scattering is re-radiation following changes of internal energy levels within individual molecules. Analysis of light scattering thus has the capability to provide physical information at many different levels.

Differential absorption, fluorescence and surface heating/vaporization are generally specific to the target and may be used with an appropriate choice of laser wavelengths to provide chemical-type information.

The available laser beam sizes and speckle effects of laser radiation are of course determined by well-established principles of optical propagation. In lidar work, the outgoing laser beam will usually be set up in the lowest order optical mode (TEMoo) with Gaussian (bell-shape) intensity distribution. At a distance z from the *minimum* beam waist diameter d_0, the beam size is given by:

$$D^2(z) = d_0^2[1 + (4\lambda z/\pi d_0^2)^2] \tag{C3.2.1}$$

This minimum beam size, d_0, occurs where the beam is collimated (i.e. where the wave front is plane). As shown in figure C3.2.1, in lidar operation, the laser may be transmitted in either one of two forms: as a focused beam on a nearby target (usually up to ~1 km) or as a collimated beam onto a distant target. In either case, from equation C3.2.1, if $0 \leqslant d(z)$ then

$$d(z) \approx \lambda_z/\pi_0 \tag{C3.2.2}$$

and the angular divergence of the beam is given by

$$d(z)/z \approx 4\lambda/\pi_0 = 1.27(\lambda/0) \tag{C3.2.3}$$

Table C3.2.3. Some basic principles and phenomena for laser sensing.

Light scattering: elastic, specular, diffuse, Mie, Rayleigh, Brillouin, Raman
Differential absorption
Fluorescence
Surface heating and vaporization
Light propagation (diffraction and speckle)
Doppler effect
Interference (interferometry; light mixing/beating/heterodyning)
Timing principle

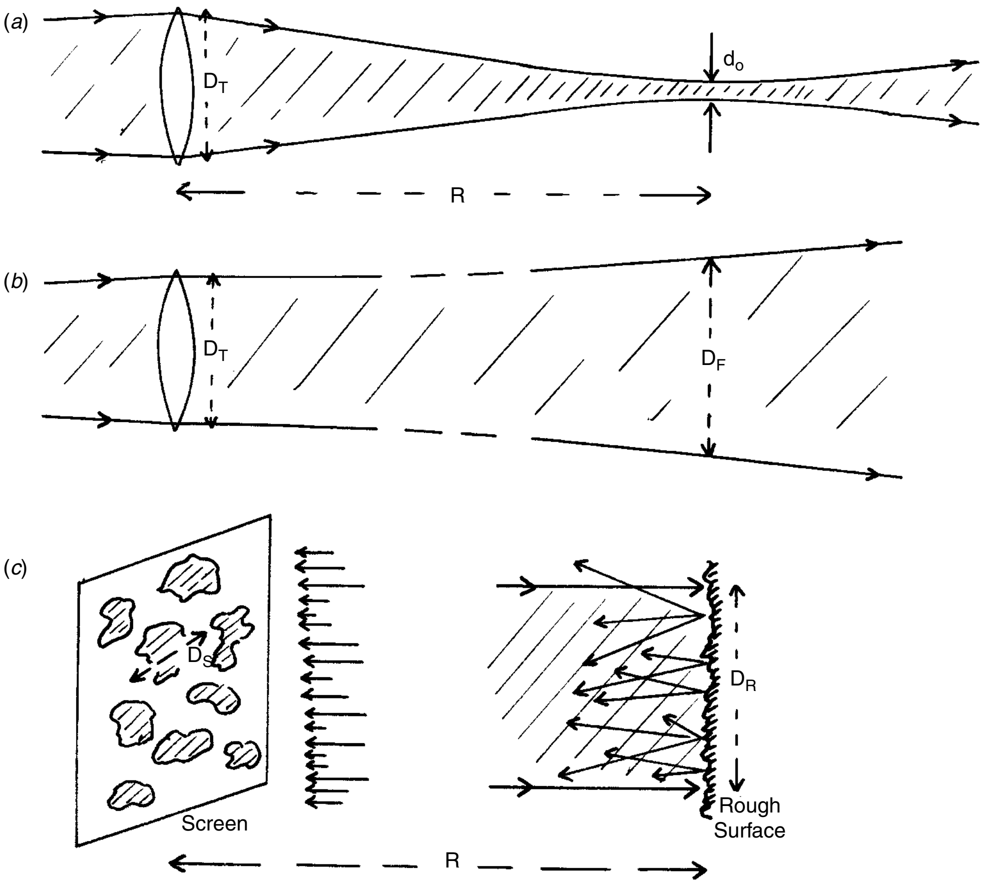


Figure C3.2.1. Illustration of the focusing of laser beams and laser speckle. (*a*) Converging beam of diameter D_T transmitted from a lidar telescope and focusing to a diameter 0 at range R. (*b*) Collimated beam (with plane wavefront) D_T propagating to a diameter D_F at a distant range. (*c*) Laser speckles (of typical size $\sim D_s$) due to a scattering from a rough surface illuminated with a laser beam of size D_R.

From these expressions, it is easy to show that a 10 cm diameter beam (D_T in figure C3.2.1*a*) at $\lambda = 10.6\,\mu$m may be brought to a focus of $0 = 1.36$ cm at a range R of $z = 100$ m. Conversely, for a collimated beam at the transmitter with $D_T = 0 = 10$ cm (in figure C3.2.1*b*) the beam size D_F at 10 km range and $\lambda = 10.6\,\mu$m would be 1.36 m. These values of beam diameter reduce very sharply with shorter wavelength λ (note the λ term in the numerator of equation C3.2.2). However, the refractive effects of atmospheric turbulence are much greater at the shorter wavelength. These introduce optical aberration in the light path so that, in the lower atmosphere at least, it is difficult to focus the shorter-wavelength beams any more sharply than longer wavelengths (see section C3.2.3.4 following). This, of course, no longer holds true high in the atmosphere or in space.

Figure C3.2.1c illustrates a laser beam of radius D_R illuminating a rough surface that is shown as an assembly of small scatterers, which reflect the beam in all directions. Typically, at a distance R, the interference of the light from all these many scatterers will produce a random speckle pattern.

The characteristic size D_s of these speckle blobs is given by a similar expression to equation C3.2.2 with

$$D_s \approx \lambda \cdot R/D_R \tag{C3.2.4}$$

Thus, as the beam illuminating the target gets bigger, the speckle size decreases. The fluctuations of speckles, and their temporal and statistical characteristics for moving targets, are a complex and fascinating topic and will be touched on in various sections of this report.

The next item in table C3.2.3, the Doppler effect, is the well-known change in frequency of a wave from a moving object (as typified, for example, from a passing train whistle or a police siren). Exactly the same thing happens in the light radiation scattered from a moving target. In back scattering (i.e. that light scattered back towards the transmitter) the Doppler shift f_D is given by

$$f_D = 2V_R/\lambda \tag{C3.2.5}$$

This is illustrated in figure C3.2.2*a* where V_R, equal to $V\cos\theta$, is the radial line-of-sight component of motion of the scattering target.

Laser radiation in the visible and near-visible region is, of course, of a very high frequency that cannot be followed by available detectors. Green light of $\lambda = 0.5\,\mu\text{m}$, for example, is of frequency $f = 6 \times 10^{14}$ Hz or 600 THz and from equation C3.2.5 the Doppler shift at $\lambda = 0.5\,\mu\text{m}$ due to a moving target with $V_R \equiv 1\,\text{m}\,\text{s}^{-1}$ is only 4×10^6 Hz or 4 Mhz. Such a small shift may be measured with advanced classical spectroscopic techniques of high-resolution interferometry, employing, for example, Fabry–Perot etalon filters. However, the interference principle may also be employed to measure this Doppler shift as added to the much higher 'carrier' frequency of the laser f_L. Thus, if the shifted and unshifted beams in figure C3.2.2*b* are superimposed at a detector, the electromagnetic fields of the local oscillator beam at frequency f_L and the scattered beam (from a moving target) of frequency $(f_L + f_D)$ will beat or heterodyne together. In consequence, the detector will provide an oscillating electrical signal i_S at the difference frequency $(f_L + f_D) - (f_L) = f_D$, as described in the following section.

Finally, we consider the timing principle as indicated in figure C3.2.2*c*. If a laser beam is transmitted as a very short, sharp pulse of length Δt_p, the measured time-of-flight t_f to and from a target gives a measure of the range R and range resolution ΔR as

$$R = ct_f/2 \tag{C3.2.6a}$$

$$\Delta R = c\Delta t_p/2 \tag{C3.2.6b}$$

where c is the velocity of light ($\sim 3 \times 10^8\,\text{ms}^{-1}$). Thus, a 10 ns pulse system will readily provide a range resolution of 1.5 m, which may be adequate for many purposes; shorter pulses may be readily employed to provide centimetric precision or better.

C3.2.2.2 Basic techniques: incoherent direct-detection and coherent light-beating

For many applications, the scattered light from the target may be directed to a detector with no more than a simple optical filter to cut down most of the background light. However, for more advanced applications, as outlined briefly above and shown schematically in figure C3.2.3, there are two basic lidar techniques for detection and spectral analysis of the scattered light field:

- Direct detection or optical/frequency domain spectroscopy in which the light field is operated on by optical elements prior to detection.
- Post-detection or time domain spectroscopy in which analysis is conducted after the light field is detected, often in conjunction with a coherent heterodyne local oscillator beam.

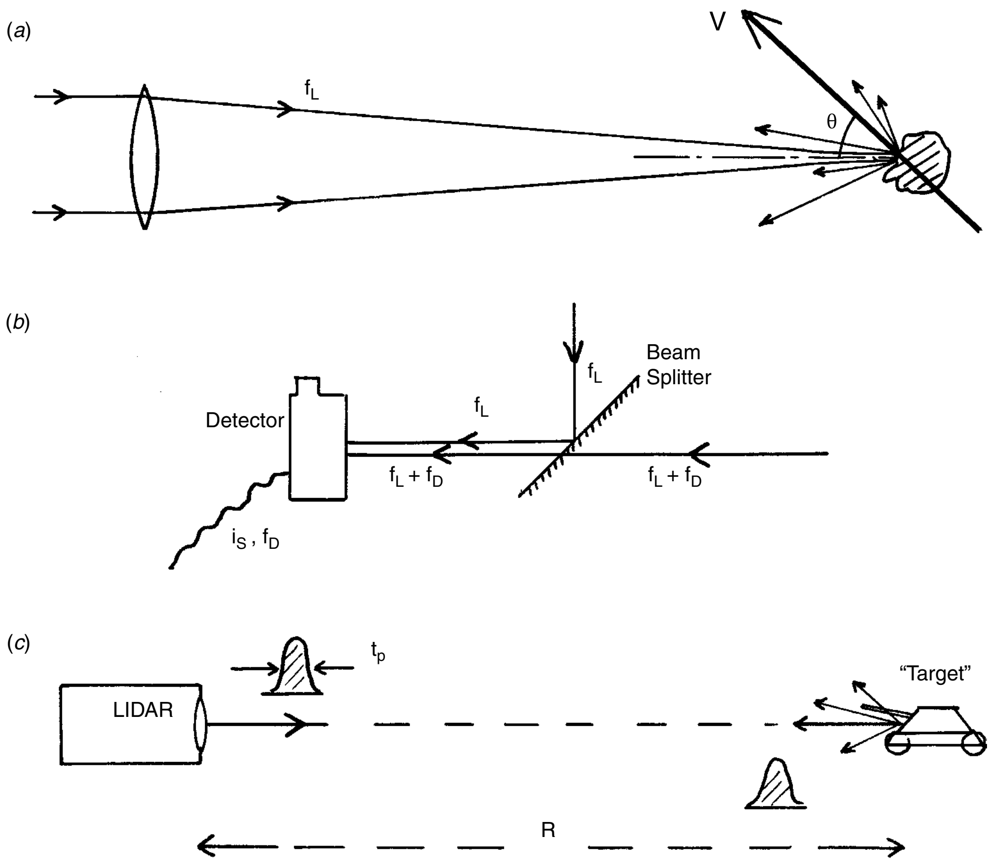
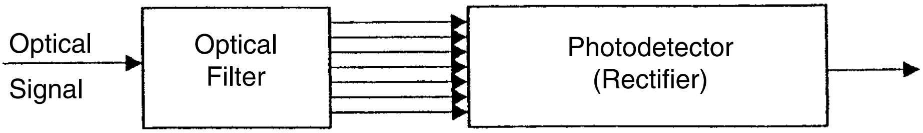


Figure C3.2.2. (*a*) Schematic of the Doppler frequency shift in light backscattered from a moving object. As shown, the line-of-sight velocity component is $V \cos \theta$ and thus the Doppler shift is $f_D = 2V \cos \theta/\lambda$. (*b*) Illustration of light beating or heterodyning by mixing the two beams at the surface of a detector which provides an oscillating electrical signal i_s at the difference frequency f_D. (*c*) Schematic of a simple time-of-flight laser rangefinder. If the measured time delay to the target and back is t_f the range R is $ct_f/2$.

The two techniques are thus different at a very fundamental level and have been extensively discussed (see, e.g. Refs. [2, 3]). In the older direct-detection techniques, the beam may be operated on by various classical interferometric devices (two-beam: Michelson, Mach–Zehnder, etc; multiple-beam: Fizeau, Fabry–Perot, Lummer–Gehrke plate, echelle, grating, etc) to form a visibility curve or spectrum from which spectral information may be derived. In post-detection, coherent heterodyne methods, the spectrum is formed (or equivalent information derived) by manipulation of the electrical signal as it emerges serially in time from the detector element.

For coherent operation with full mixing efficiency, very good control of the local oscillator beam and scattered signal beam is required, with very precise matching of the wavefronts and optic axes. A fundamental characteristic of such an arrangement is that, in effect, the local oscillator beam selects just a single mode of the scattered radiation field. This ensures that the full speckle characteristics of the scattered radiation (amplitude, frequency, phase), with all the dynamic and fluctuation parameters, are fully preserved in the measured electrical signal (in contrast with direct detection where these properties

Direct Detection



Heterodyne Detection

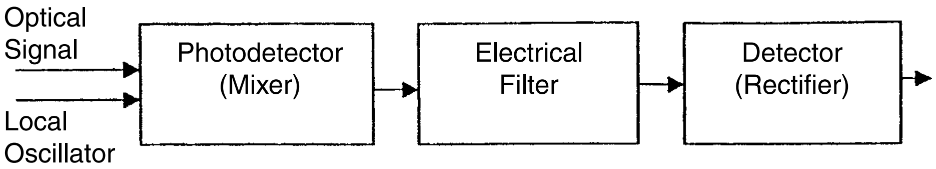


Figure C3.2.3. Schematic of the two basic forms of laser radar showing the flow diagram of optical to electrical processing. Direct detection lidar in which the scattered light is collected by the receiver and passed directly via an optical filter to the detector. At high resolution, interferometric stability will be required in the optical filter. Coherent heterodyne lidar in which the scattered light is mixed with an optical local oscillator beam derived from the original laser as in figure C3.2.2*b*. In this case, very precise, interferometric control and matching optic axes and wavefronts of the two beams are required as they beat together at the surface of the detector.

are necessarily smeared out by integration over a finite optical aperture). In essence, the coherent technique provides a measure of the vector sum of the amplitude of electromagnetic field, $\check{E}_s$, over the detector, whereas, in direct detection of the scalar sum of intensities, $|E_s^2|$ is formed. The coherent beating technique thus provides an important and perhaps little appreciated spectroscopic tool; probability distributions, higher moments, autocorrelation functions, structure functions and fractal character may be investigated for different fields (see, e.g. Ref. [4]) and there is wide application to many aspects and phenomena of laser physics.

Not surprisingly, the information transfer by these two different forms of operation on an electromagnetic field is very different. The most significant factor is the rate of photons detected within a single coherence area (an étendue of λ^2) per unit band pass and, most particularly, the number of detections from one cell of phase space, equivalent to the number of counts per coherence time and usually called the photon degeneracy parameter δ. In coherent time-domain techniques operating on a single optical mode, the available étendue or light grasp is limited to λ^2. On the other hand, in direct-detection, interferometric methods, the optical filter acts directly on the light field and one measures the direct current component of the signal collected in a beam of étendue U, which can be many times larger than λ^2.

A summary comparison of the two techniques is given in table C3.2.4. As will be seen in the following sections, the basic choice of technique will be dictated by the problem in hand and the most appropriate laser source for that problem. In broad terms, the large photon energy at shorter wavelengths means that detectors should be comparatively free of noise, so direct detection techniques may be preferred for $\lambda \preccurlyeq 1.5\ \mu\text{m}$. At longer wavelengths with heterodyne techniques, shot noise in the local oscillator beam may be used to dominate thermal noise in the detector to provide quantum-limited

Table C3.2.4. Comparison of direct-detection and coherent techiques.

Direct Detection	Coherent Heterodyne
Scattered radiation optically manipulated before being passed to detector	Scattered radiation mixed (heterodyned) with coherent local oscillator beam
Measures scattered intensity in selected frequency intervals	Full phase, frequency and amplitude information available
Light may be collected in many optical modes	Light collected in a single optical mode
Some relaxation of laser characteristics may be available (depending on frequency resolution required)	Laser requires to be of high spatial and spectral purity
Relatively simple optical arrangement of transmitter and receiver but interferometric stability required in analyser	Very precise, stable, optical arrangement required to maintain wavefront matching of scattered and local oscillator beams
Quasi-noise free detection at shorter wavelengths; additional noise source at longer wavelengths	Detector noise at longer wavelengths may be advantageously dominated by the local oscillator shot noise

detection of the signal beam. A more detailed evaluation of signal-to-noise ratio and measurement accuracy is provided in section C3.2.3.3 following.

C3.2.3 Lidar technology and systems

C3.2.3.1 The building blocks of a lidar system

The basic elements of a laser radar are indicated in figure C3.2.4. In the design of a system for any given task, there are many aspects to be considered and choices to be made. Immediately obvious factors include the nature of the target and the information required from it, the environment and type of platform (ground, air or space) from which the lidar must operate and not least the available financial budget. Some of the vast range of technical considerations and options are shown in table C3.2.5. This list serves to establish both the intellectual challenge of laser radar and the diversity it offers of basic science, advanced technology and field measurement. In the present work, it is obviously not appropriate to consider all the various building blocks shown in figure C3.2.4 and table C3.2.5 in any detail. It is, however, worth outlining a number of issues. These include: (1) propagation in the atmosphere, the choice of lasers and the relative merits of working at various wavelengths; (2) basic questions of signal to noise and calibration of lidars; and (3) some consideration of range resolution and optical and system design issues for operation from different platforms.

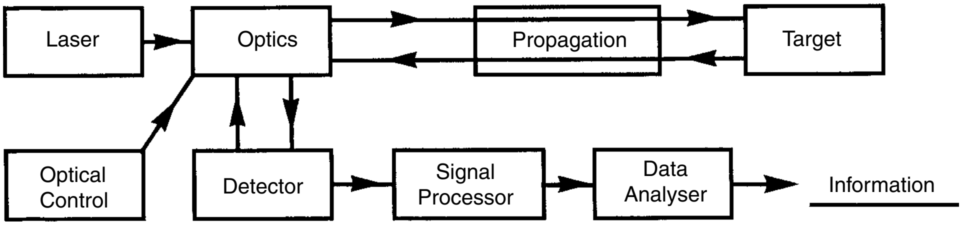


Figure C3.2.4. The basic components of a laser radar. See text and table C3.2.5 for an outline of the many technical considerations and options for the various components.

Table C3.2.5. Some technical considerations and options in the design of a coherent laser radar (see also figure C3.2.4).

Laser	Power level, pulsed/cw, wavelength, line stability, coherence, lifetime, mass, volume, power supply, cooling, etc
Optics	Size, uniaxal/biaxal, optical/mechanical precision, field of view, temperature effects, long-term stability, etc
Optical control	Scan patterns, speed, range discrimination, aiming, switchability, etc
Propagation	Attenuation, aberrations—turbulence, absorption, etc
Target	Scattering, signal strength, range, Doppler shift, signal fluctuations and statistics, speckle, coherence time, etc
Detector	Bandwidth, sensitivity, noise characteristics, cooling, etc
Signal processor	Frequency, integration times, repetition rate, thresholds, missed signals, false alarms, etc
Data analyser	Parameters, integration times, repetition rate, thresholds, missed signals, false alarms, etc
Information	Range, velocity, bearing, elevation, character, signal strength, fluctuations, error rates, etc

C3.2.3.2 Atmospheric transmission and choice of laser

For field operation at useful ranges, one generally requires a laser with good transmission through the atmosphere (although it may be noted that, for some operations, a lidar with limited sensing out to a few tens of metres, and which thereafter is rapidly absorbed, may be valuable). Figure B9.2 shows a low-resolution spectrum of transmission in the atmosphere near ground level in clear weather. Only certain regions in this spectrum (known as 'atmospheric windows') can be used for remote sensing over extended ranges. As is obvious in Figure B9.2, the absorption due to molecular components, primarily water vapour and carbon dioxide, in the atmosphere establishes three primary bands of good transmission: from the near-ultra violet at ~0.3 μm through the visible region and up to ~2.5 μm; 3–5 μm; and the longer-wavelength 8–13 μm region. Table C3.2.6 shows a brief list of lasers that may be used. At the shorter wavelengths, a vast range of lasers is available, including gas lasers (e.g. argon ion Ar^+, helium–neon He–Ne, etc), solid state (e.g. ruby, neodymium–yttrium aluminium garnet, Nd-YAG, etc) and semi-conductor and fibre amplifier lasers.

The 3–5 μm range is less well served with CO and various He–Ne–Xe lasers and OPOs currently offering rather low efficiency in conversion of electrical to optical laser power (so called 'wall-plug' efficiency). In the 8–12 μm band, carbon dioxide CO_2 lasers are widely used and offer both high efficiency (often $>10\%$) and a very wide tuning range over many possible molecular lasing transitions.

Selection of the laser type and most effective lidar for a given remote-sensing problem are obviously determined by a vast range of factors, some of which are indicated in table C3.2.5 and further discussed in the following section. Any decision is often a matter of best engineering judgement and not susceptible

Table C3.2.6. Some of the lasers that may be employed for lidar work in the visible and near infrared.

Gas lasers:	He–Ne 0.63 μm; Ar^+ 0.45–0.55 μm; CO, He–Ne–Xe 3–5 μm; CO_2 9–12 μm
Solid state:	Nd-YAG—1.06 μm (and frequency doubled 0.53 μm and frequency tripled 0.35 μm); ruby 0.65 μm; Tm, Ho: YAG: YLF 2–2.2 μm
Diode lasers:	0.8–1.6 μm with power augmented in optical fibre amplifiers; distributed feedback (DFB) lasers 1.55 μm
OPOs:	Optical parametric oscillators, tunable 3–5 μm

of scientific ‘proof’. Over the years, considerable controversy has sparked between advocates of different lasers and systems; such arguments have usually generated rather more heat than light and on occasions have delayed progress.

C3.2.3.3 Signal strength, signal-to-noise and lidar calibration

The signal return power, collected by a laser radar from a large, solid, hard target (bigger than the laser beam), can be written as

$$P_{\mathrm{R}} = P_{\mathrm{T}} \cdot \varepsilon [A/\pi R^2] \cdot T \cdot e^{-2\alpha R} \tag{C3.2.7}$$

where P_{T} is the transmitted power, ε is the target reflectivity, A is the effective collection area of the lidar receiver and R is the range to the target from the lidar. Note that, if the target is small compared with the beam, the range dependence becomes $1/R^4$. The atmospheric absorption coefficient for the operating wavelength is α and T is the transmission of the lidar optics. For a direct detection lidar (figure C3.2.3*a*), the effective collection aperture A_{dir} is given by the area of the lidar-receiving telescope (and as dictated by the available étendue U of the analysing interferometer). However, for coherent lidar (figure C3.2.3*b*), the situation is somewhat more complex, as discussed in section C3.2.2.2. To a good approximation, with a well-adjusted system, the effective area of the collection aperture for a heterodyne system can be written as

$$A_{\mathrm{het}} = [1/A_{\mathrm{SP}} + 1/A_{\mathrm{LO}} + 1/A_{\mathrm{AT}}]^{-1} \tag{C3.2.8}$$

where A_{SP} is the average area of a speckle element (figure C3.2.1*c*) in the return field. The influence of the local oscillator and its spatial profile is incorporated by the A_{LO} term. Atmospheric turbulence is incorporated by the A_{AT} coherence area term. Turbulence effects will dominate, and the return signal will be greatly reduced, if A_{AT} is smaller than the combination of the unperturbed speckle and local oscillator areas A_{SP} and A_{LO}. Physically, this means that, with reference to figure C3.2.1*c*, the speckles have been broken up by the turbulence-induced refractive index fluctuations and are much reduced in size.

In a lidar experiment, the accuracy with which a parameter may be measured will be governed by the signal-to-noise ratio (SNR). This in turn is fundamentally constrained by the quantum nature of light and the number of photons in the signal. Thus, for an experiment lasting a time t_{e} with a signal power P_{S} and effective quantum efficiency of the detector of η, the mean number N_{S} of signal photons detected (each of energy $h\upsilon$ where h is Planck’s constant and the angular frequency of the light υ is equal to $2\pi c/\lambda$) will be given by

$$N_{\mathrm{S}} = \eta (P_{\mathrm{S}}/h\upsilon) t_{\mathrm{e}} \tag{C3.2.9}$$

The SNR relations for direct and coherent detection are, in fact, rather different. With reduced spectral discrimination in direct detection, there is the possibility of collecting background light N_{b} (e.g. daylight) not related to the signal, and also a noise contribution N_{d} from the detector itself so that the average measured signal $\langle N_{\mathrm{m}} \rangle$ is given by

$$\langle N_{\mathrm{m}} \rangle = \langle N_{\mathrm{S}} + N_{\mathrm{b}} + N_{\mathrm{d}} \rangle \tag{C3.2.10}$$

From this, it might erroneously be supposed that the SNR would be given by $N_{\mathrm{S}}/(N_{\mathrm{b}}+N_{\mathrm{d}})$. This, however, provides no indication of the accuracy with which the signal can be measured, for which we need the fluctuations in these quantities. The magnitude of the root variance N_{S} is, in fact, for commonly met statistics, given by $N_{\mathrm{S}}^{1/2}$, with similar expression for background and detector noise. Thus, measurements of detector output, with and without signal present, provide an $\mathrm{SNR}_{\mathrm{dir}}$ given by:

$$\mathrm{SNR}_{\mathrm{dir}} = (\text{measured signal})/\{\text{sum of variances}\}^{1/2} = (N_\mathrm{s})/[(N_\mathrm{s}+N_\mathrm{b}+N_\mathrm{d})+(N_\mathrm{d}+N_\mathrm{b})]^{1/2}$$
$$= (\eta^{1/2}\mathrm{P}_\mathrm{s}t_\mathrm{e}^{1/2})/(h\upsilon)^{1/2}[(P_\mathrm{s}+P_\mathrm{b}+P_\mathrm{d})+(P_\mathrm{d}+P_\mathrm{b})]^{1/2} \qquad \text{(C3.2.11)}$$

In the case where $P_\mathrm{S} \gg P_\mathrm{b}$, P_d the SNR reduces to $N_\mathrm{S}^{1/2}$. In this case, the 'noise' is truly 'noise-in-signal', as governed by the quantum nature of light detection.

In coherent heterodyne detection, due to the mixing of the local oscillator light field E_LO and the signal light field E_S, the detector provides an electrical signal i_s proportional to their product so that

$$i_\mathrm{S} \propto E_\mathrm{LO}\,E_\mathrm{S} \qquad \text{(C3.2.12)}$$

Analysis shows that, in a well-adjusted system, the coherent $\mathrm{SNR}_{\mathrm{coh}}$ in the power spectrum of the signal is given by

$$\mathrm{SNR}_\mathrm{c} = \eta P_\mathrm{s}/h\upsilon B = N_\mathrm{s}/B \qquad \text{(C3.2.13)}$$

where η is the effective quantum efficiency of the detector and B is the operational frequency bandwidth for the lidar system. It is supposed that the local oscillator has been arranged so that its own photon shot-noise is the dominant term above any contribution from even an inherently noisy detector. The bandwidth term B requires careful consideration. In the lidar signal processing, the instrumental bandwidth B_I requires to be well matched to the frequency band B_S over which the signal is spread. If B_I is set much greater than B_S, the $\mathrm{SNR}_{\mathrm{coh}}$ is unnecessarily diminished. If it is set much smaller, not all the signal photons are effectively utilized.

The precision with which frequency measurements (e.g. Doppler shifts) may be made can be analysed in terms of the Cramer–Rao relationship. For direct detection, interferometric spectroscopy, the Cramer–Rao relationship gives the limiting (standard deviation) accuracy δ*W*, with which the centre of a Gaussian spectral line profile can be found, as

$$\delta W \geq 0.425 W_\mathrm{G} N_\mathrm{S}^{-1/2} \qquad \text{(C3.2.14)}$$

where N_S is the total number of photocounts distributed across such a line profile of width W_G (half width at height $e^{-\frac{1}{2}}$). This is illustrated in figure C3.2.5 by the lower line of slope −1/2, showing attainable accuracy versus total photocount. These total photocounts may be accumulated by summation over many individual laser pulses and, as noted, across a collection aperture containing many optical modes (provided they lie within the allowable étendue of the interferometer).

For coherent heterodyne spectroscopy, the situation is very different (see, e.g. Ref. [2]). The simple Cramer–Rao relationship of equation C3.2.14 needs to be multiplied by a complex term containing the photocount degeneracy δ. The effect of this is manifest in the upper curves of figure C3.2.5 for accumulation of $n = 1$ and $n = 100$ laser pulses. As outlined in the previous section, the light signal must be collected within a single optical mode and for a low photocount the slope (for degeneracy $\delta < 3.3$) is close to -1 and the accuracy is best improved by increasing the laser energy per pulse. Only for a higher photocount ($\delta > 3.3$, deriving from stronger scattering) is the accuracy best improved by accumulating over a number of pulses. There is thus a very clear difference in the design considerations for the two classes of lidars. For direct detection, the available laser energy may be spread without undue penalty over many pulses—the total effective photocount and accuracy depend only on the total energy transmitted (and proportionately scattered). For coherent heterodyne spectroscopy, it is essential (at lower scattering levels) to make the energy per pulse as large as possible to ensure that δ is as large as possible.

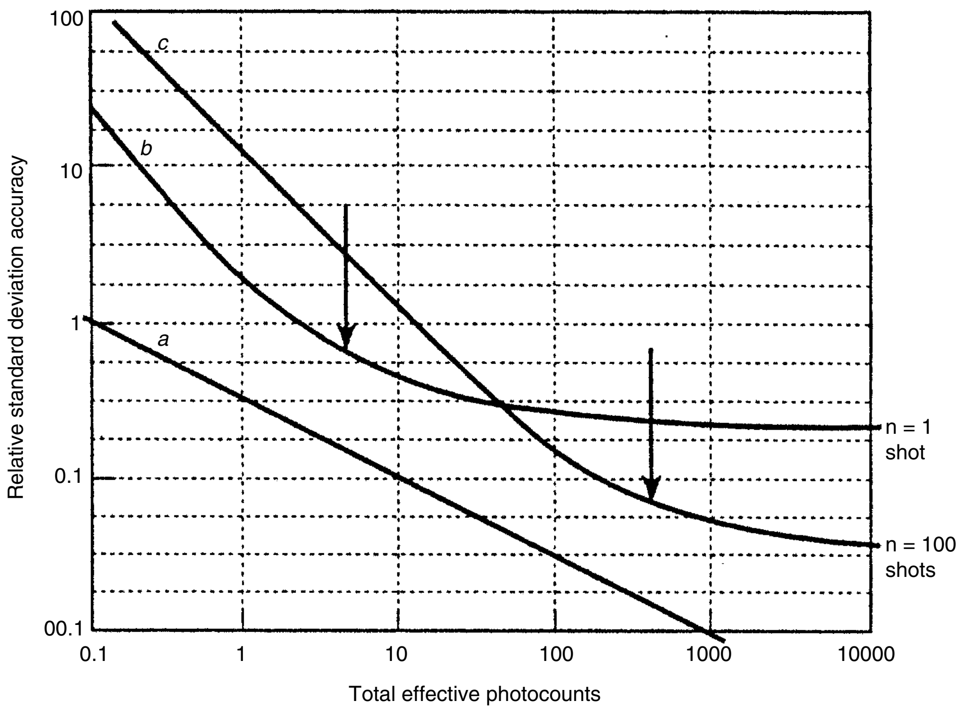


Figure C3.2.5. Schematic of the standard deviation of Doppler frequency estimates as a function of the total effective photocount on log–log scales. Line (*a*) shows the direct detection Cramer–Rao limit with slope $-1/2$. Curve (*b*) is the heterodyne Cramer–Rao for $n = 1$ shot (i.e. pulses); note the slope of -1 at low photocount and tending to zero and saturation at high photocount. The arrow indicates the region of closest approach with photocount degeneracy $\delta \approx 3.3$. Curve (*c*) is the heterodyne Cramer–Rao for $n = 100$ shot (i.e. pulses): this is a translation of curve (*b*) by a factor of 100 in effective photocount and a factor on $(n)^{1/2} = 10$ in standard deviation accuracy.

Further examination of signal statistics, processing, etc, is beyond the scope of the present chapter. The subject rapidly becomes extremely complex with many highly specialized techniques. It is, however, worth considering the topic of calibration and performance (see, e.g. Refs. [5–8]). Equations C3.2.7 and following relate the lidar signal to the lidar parameters. If any of these latter are in any way defective, e.g. poor optical transmission T, inefficient detector η, reduced laser power P_T, etc, the system performance will be impaired. It is, in fact, often remarkably difficult to prove by absolute calibration test of signal strength that a lidar system (particularly a coherent system) is working to its full potential. The principle of calibration is fairly straightforward; given a test target of well-known scattering characteristics (with reliable ε in equation C3.2.7) a $\mathrm{SNR}_{\mathrm{calc}}$ is *calculated* for the lidar from all its individual, precisely-measured, parameters (e.g. optical transmission of lenses, beam splitters detector quantum efficiency, etc) as inserted into the lidar equations. This value is then compared with that $\mathrm{SNR}_{\mathrm{obs}}$ which is actually observed at the output of the lidar signal processing. Any major discrepancy requires explanation and a search for defective elements, alignment and signal processing components. It is of course extremely important to conduct such an exercise; a 3 dB (factor of 2) loss in $\mathrm{SNR}_{\mathrm{obs}}$ for example would either require in compensation a doubling of laser power or provide a range performance reduced by at least 30%.

C3.2.3.4 *Range resolution, optical and system design issues*

For pulsed lidars, the range and range resolution are given by equations C3.2.6*a* and *b* and the selection of time delay t_f, length of laser pulse Δt_p and the length of the timing interval Δt_g (often called the 'range gate') in the signal-processing system. Pulse length and range gate are usually set to be rather similar for most effective operation. As noted in section C3.2.2.1, a 10 ns pulse gives a range resolution of 1.5 m, which may provide adequate accuracy for range measurement on a solid distant object. For an extended target, such as the atmosphere, measurements over successive volumes along the laser beam may be required, in which case, for example, a 1 μs pulse (and matched signal-processing range gate) would give ~ 150 m resolution. In this case, measurements up to 15 km would require 100 equal range gates, each of equivalent length 150 m in the processing.

For direct detection, the signal strength has a range dependence given by equation C3.2.7 and reduces as R^{-2} (supposing that all the transmitted energy is incident on the 'target'). The corresponding range dependence for the coherent heterodyne case was first evaluated in 1971 by Sonnenschein and Horrigan [9] (see also Ref. [10]); it is implicit in equation C3.2.8 but can be expressed more directly as

$$S(R,F) \propto [R^2\{1 + (A/\lambda)^2(1/\mathrm{R} - 1/F)^2\}]^{-1} \qquad \text{(C3.2.15)}$$

where F is the range at which the lidar beam is set to focus and A is the effective area of the telescope radius (~A_{LO} in equation C3.2.8). This expression has some interesting implications; for a target at a given range R, the maximum signal is attained by focusing at the target and setting F equal to R. Thus, for solid targets (so-called 'hard' targets with scattering from a solid surface) expected at various long ranges, the lidar would be set with the outgoing beam almost collimated; the R^2 term dominates in the bracket on the left-hand side of equation C3.2.14 and the range dependence (R^{-2}) is the same as for direct detection. Similarly, for a short range, the signal strength will peak up at ranges R close to F. This is illustrated in figure C3.2.6, where a representative value of $(A/\lambda)^2$ equal to $3 \times 10^6\,\mathrm{m}^2$ has been used. This establishes the possibility of range selectivity with a continuous wave laser beam and is particularly useful for measurements on an extended or diffuse source of scatterers, such as small particles in the atmosphere. In this case, the measured coherent signal derives most strongly from the focal region of the beam. This comes about because the scattered radiation from this region is best matched in wave front to the local oscillator beam in the lidar. In consequence, it beats most efficiently with the local oscillator and provides the strongest signal. The curves shown in figure C3.2.6 are representative of what may be achieved with a 10.6 μm coherent lidar with a 30 cm diameter transmitting/receiving telescope (with the outgoing laser beam and equivalent A_{LO} of best diameter ~ 16 cm). The central probe volume is commonly defined as the region within which the signal is within 3 dB of its peak position: its half-length is written as $\Delta R_{3\,\mathrm{dB}}$. Some manipulation of equation C3.2.15 with the approximations $F \approx \gg \Delta R$ gives:

$$\Delta R_{3\,\mathrm{dB}} = \pm F^2(\lambda/A) \qquad \text{(C3.2.16)}$$

Thus, at a focal setting of $F = 100\,\mathrm{m}$, a spatial resolution of about ±6 m is produced. For such diffuse, extended targets, the total observed signal S will be determined by the integral over an appropriate range interval R_1 to R_2 given by

$$S \propto \int_{R_1}^{R_2} P(R)S(R,F)\delta R \qquad \text{(C3.2.17)}$$

where $P(R)$ is the strength of scattering at range R. In a generally clear uniform atmosphere, $P(R)$ over short ranges may be considered as constant, but obviously this is a very crude approximation in conditions of low cloud, layered fog or smoke [11].

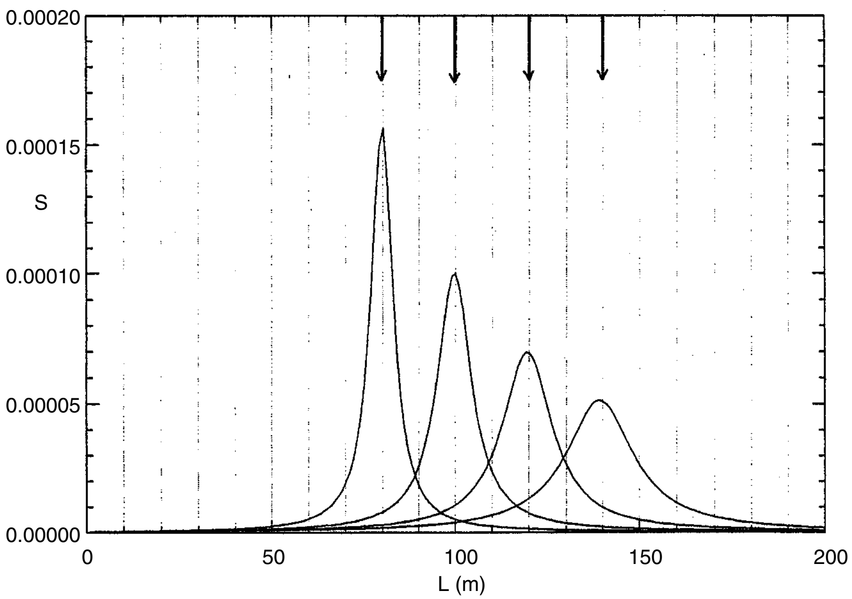


Figure C3.2.6. The range sensitivity for a coherent lidar as given by equation C3.2.14 with the $(A/\lambda)^2$ terms equal to $3 \times 10^6\,\mathrm{m}^2$, and for four values of the focus $F = 80$, 100, 120 and 140 m. Note how the sensitivity peaks up very strongly around the focal range F. This permits the possibility of useful range resolution with a continuous wave laser for measurements on an extended diffuse target such as aerosols in the atmosphere.

Brief consideration of these equations indicates some of the complexity of comparing coherent lidar performance at different wavelengths. Thus, at shorter wavelength, the focal depth will be smaller and the spatial resolution will be greater; the total signal S is correspondingly reduced for the smaller ΔR. However, at shorter wavelength, the maximum useful size for the effective telescope diameter, D, is limited by refractive turbulence in the atmosphere (which destroys lateral coherence across the beam, thus reducing A_{SP}; see section C3.2.2) and has an impact approximately proportional to $\lambda^{-6/5}$. Thus, if atmospheric conditions permit, an effective telescope diameter of ~50 cm at 10.6 μm wavelength would be reduced by a factor $(10.6/2.06)^{6/5}$ [12] (for an extensive discussion with many references, see also Ref. [13]) at 2.06 μm to only ~7.0 cm. Such a reduced collection aperture would provide both a weaker signal and poorer spatial resolution than the admittedly larger and more expensive telescope that could be used in these circumstances at the longer wavelength. Such wavelength comparisons may be extended to many other aspects of signal collection, processing and measurement precision. Thus, in equation C3.2.13, the instrumental bandwidth required to cover a given velocity interval is proportional to λ^{-1} (see equation C3.2.5 for the Doppler shift) and altogether with the larger photon energy $h\upsilon$ gives

$$\mathrm{SNR_c} \propto \eta P_{\mathrm{s}}.\lambda^2 \tag{C3.2.18}$$

For a given laser power, the scattering P_{s} from most surfaces and aerosol particles increases at shorter wavelength. To give a signal-to-noise advantage at shorter wavelength, this rate of increase obviously needs to be greater than λ^{-2} to overcome the longer wavelength advantage of equation C3.2.18.

Another point of comparison is the choice of pulse length Δt_{p}. Fourier transform analysis of a short wave train gives a limiting frequency resolution $\Delta f_{\mathrm{T}} \sim (\Delta t_{\mathrm{p}})^{-1}$. Thus, a 1 μ pulse gives a line width of ~1 MHz. Comparing this with equation C3.2.5, such a line width would provide a velocity resolution ΔV given by

$$\Delta V = \Delta f(\lambda/2) = \lambda(2\Delta t_{\mathrm{p}})^{-1} \tag{C3.2.19}$$

Thus, at 2 μm, a pulse length of 1 μs would give a Fourier transform velocity resolution of ~1 m s^{-1}, which increases to ~5 m s^{-1} at 10 μm. This potentially provides a significant advantage to shorter wavelengths and also offers the possibility of achieving greater spatial resolution (with shorter pulses).

Further discussion is beyond the scope of the present article; wavelength and system comparisons are generally very difficult and rarely permit a clear cut 'winner'. Selection of a lidar system for a given task usually requires consideration of many issues, often practical and logistical, as well as the technical options outlined in table C3.2.5. As an example, the overriding constraints for an aircraft or military installation is likely to be the maximum optical aperture that can be accommodated in the structure, together with strict limits on mass, volume and available power budget.

C3.2.4 Lidar and atmospheric sensing

C3.2.4.1 Introduction: atmospheric parameters and measurement

Some of the first demonstrations of lidar in the 1970s were measurements of wind and cloud. The light scattered from a laser beam in a clear atmosphere is primarily from air molecules (Rayleigh scattering) and from small airborne particles (dust etc), usually referred to as aerosol scattering. The strength of molecular scattering is approximately proportional to λ^{-4}, whereas the wavelength dependence of aerosol scattering typically lies between λ^{-1} and λ^{-2}. Thus, molecular scattering is much stronger at shorter wavelengths (hence the blueness of the sky) and also increases relative to aerosol scattering. Perhaps the most obvious common example of the latter is the appearance of a shaft of sunlight between clouds.

Wind and atmospheric measurements are interesting and important for many reasons: to meteorology and climatology, for example, for the information they provide on the behaviour of the atmosphere, to local studies of shear and turbulence, including the planetary boundary layer and practical measurements for shipping, aviation and commerce. All these impact the four primary parameters of interest to the atmospheric physicist shown in table C3.2.7.

Lidar studies make a contribution to all these areas and this is likely to increase in the future with advanced instruments operating remotely and from space.

Table C3.2.7. Primary atmospheric parameters.

Structure	Temperature, density and pressure as a function of altitude
Dynamics	Circulation and motion at all scale sizes from global through synoptic to small local scale
Radiative properties	Distribution, characteristics and structure of clouds and aerosols and their impact on solar, ground and atmospheric radiation
Chemical concentration	Vertical structure and horizontal distribution of molecular species including minor constituents such as ozone, radicals and active molecules (section C3.2.6)

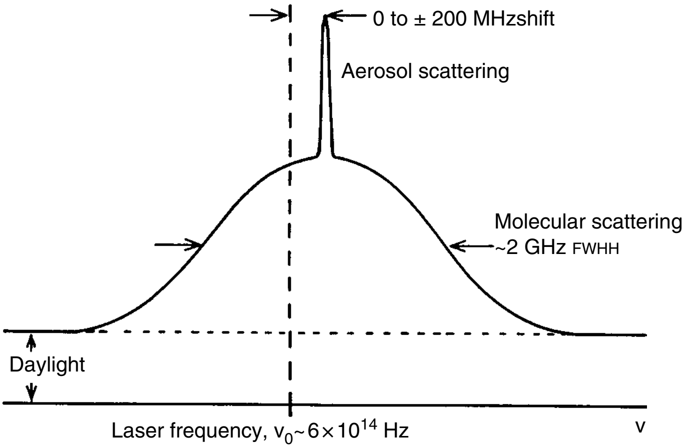


Figure C3.2.7. Schematic of the light collected by a lidar operating in the visible region with components due to aerosol scattering, molecular scattering and background daylight. The relative proportion of these varies quite strongly with wavelength and height in the atmosphere. In the troposphere (up to ~10 km) and in the infrared at 10 μm aerosol scattering is generally dominant.

C3.2.4.2 Scattering in the atmosphere

Over the past 30 years, considerable effort has been put into the measurements of scattering in the atmosphere in many different regions, seasons and wavelengths. A brief account of molecular, aerosol and cloud scattering data is presented as an essential preliminary, and a schematic of aerosol and molecular scattering is shown in figure C3.2.7. First, it is worth noting the nomenclature $\beta(\theta, \lambda, z)$ $\mathrm{m}^{-1}\mathrm{sr}^{-1}$ that is commonly employed for the strength of atmospheric scattering. This expresses the fractional amount of light that would be scattered from an incident light beam along a metre length of path (m^{-1}) into the equivalent unit radian solid angle (sr^{-1}) centred at scattering angle θ, at wavelength λ and at a height z in the atmosphere. Typically, this will often be abbreviated to $\beta(\pi)$ as a general term for backscattering. It is also worth noting various lengthy publications devoted in whole or part to scattering in the atmosphere, including those by Scorer [14], Hinkley [15] (in particular the chapters by Zuev [16], Collis and Russell [17] and Inaba [18]), and also the Conference Publications of the biennial Coherent Laser Radar Meetings and International Laser Radar Conferences.

(a) *Molecular Scattering*. Molecular scattering, in both theory and practice, has been extensively discussed over the years; see, for example, the early review of Fabelinski [19]. In the atmosphere, the strength of molecular scattering varies slowly, uniformly and predictably with altitude through the atmosphere. The exact exponent for wavelength scattering should be 4.09 instead of 4 to account for dispersion of the index of refraction of air. Taking the reference wavelength of $\lambda_0 = 1.06\,\mu\mathrm{m}$ and a reference height of $z_{\mathrm{mol}} = 8000\,\mathrm{m}$, the total backscattering contribution from the air molecules in the atmosphere at height z may be modelled by an exponentially decreasing function:

$$\beta_{\mathrm{mol}}(\pi, z, \lambda) = 10^{-7}(\lambda_0/\lambda)^{4.09}e^{-z/z\mathrm{mol}}\,\mathrm{m}^{-1}\mathrm{sr}^{-1} \tag{C3.2.20}$$

The molecular spectrum is quite different in the IR from what it is in the visible due to the contribution of collisional effects on the molecular scattering and hence the atmospheric pressure. At low densities (or pressure), molecules scatter independently, producing a single, near-Gaussian line shape, whereas at higher densities and longer wavelengths the central component, the so-called Rayleigh line, bisects the Mandlestam–Brillouin doublet due to scattering from moving density fluctuations (sound waves). Thus, a well-defined triplet structure prevails at near infrared (2 μm) and thermal infrared (10 μm) wavelengths, and appears to offer some opportunity for Doppler wind lidar (DWL) applications with coherent heterodyne techniques, as discussed by Rye [20].

Other molecular scattering phenomena include resonant scattering and Raman scattering. Both may be used for atomic and molecular species identification and concentration measurements of, for example, water vapour and ozone (and aerosol concentration) and study of atmospheric chemistry and dynamics in the atmosphere (see, e.g. Refs. [18, 21, 22] and section C3.2.6) in the atmosphere.

(b) *Aerosol Scattering*. The characteristics of aerosol scattering are very complex, depending on the chemical constitution of the particles, their size distribution, typical shape, etc. These are largely dictated by the previous history of the air mass. Contributory sources include material blown up by winds and convected from the Earth's surface (most dramatically, for example, the fine Saharan sand occasionally spread over Northern Europe and the loess soil from Central China which extends every spring as a plume into the Western Pacific). Other sources include volcanic eruption (e.g. El Chichon in Mexico in 1982 and Mt Pinatubo in the Philippines in 1991), forest fires, salt spray from the sea surface, man-made pollutants and the constant rain of interplanetary material (manifest, for example, in shooting stars). It has been calculated that the eruption of Mt Pinatubo alone injected about 10 km^3 of material in the form of fine dust into the stratosphere, with clear impact on the global climate for a few years thereafter. Typically, the particles most important for lidar scattering lie in the size range 0.1–10 μm diameter; larger particles settle out fairly quickly, while smaller ones scatter much less strongly.

The investigation of the nature, sources, transport and chemistry of atmospheric aerosol is assuming steadily greater importance in atmospheric science. Transport and dispersion models are an important component of the numerical modelling machinery applied to investigation in these areas, and are used to simulate the concentration and dilution (typically due to turbulence) of airborne material of all kinds. These activities include the modelling of air pollution and quality involving, for example, sulphur, nitrogen and photochemistry (see, e.g. Refs. [23, 24]). Increasingly sophisticated and accurate transport modelling is being used in studies of the source strength and background concentrations of radiatively active trace gases [25, 26]. Operationally, the modelling of aerosol transports features in the simulation of the spread of volcanic ash, of critical importance to aviation [27], and pollution resulting from nuclear or major chemical accidents and conflagrations. International protocols are in place for the management of such emergencies and extensive collaborative work had been carried out on the testing and intercomparison of transport and dispersion models (e.g. Ref. [28]).

Transport of airborne pathogens, and biota in general, also require suitable transport and dispersion models, which have been applied to pollution problems, and the physics of atmospheric dispersion on all motion scales from street canyons to global diffusion [29]. A second area of aerosol and gaseous transport studies of pressing importance is concerned with the evolution of climate forcing and change, in which sulphate aerosol and volcanic effluent are just two of a range of airborne substances subject to intensive investigation, typically involving general circulation models.

There have now been an enormous number of lidar measurements of aerosol scattering with ground, airborne and space platforms. The wavelength range covers the UV at ~0.35 μm to IR at 10.6 μm with direct detection and heterodyne techniques (see, for example, table C3.2.8 and table C3.2.9).

However, there are great problems in synthesizing these into a database of global perspective, taking account of different wavelengths, regions, seasons, histories, sampling, etc of the measurements. Two extensive airborne programmes conducted in the late 1980s and early 1990s provide a broad base of knowledge at some of these wavelengths. These were the SABLE and GABLE programmes (South Atlantic/Global Atmospheric Backscatter Lidar Experiments) of the USAF Geophysics Laboratory and the then UK Royal Signals and Radar Establishment (subsequently the Defence Evaluation and

Table C3.2.8. Recent papers on aerosol backscatter from lidar measurements over the range 0.35–1 μm.

Ansmann *et al* [30]	Marenco *et al* [38]
Barnes and Hofmann [31]	McCormick *et al* [39]
Browell *et al* [32,33]	Osborn *et al* [40]
Cutten *et al* [34]	Parameswaren *et al* [41]
Donovan *et al* [35]	Post *et al* [42]
Hoff *et al* [36]	Shibata *et al* [43]
Li *et al* [37]	Spinhirne *et al* [44]

Table C3.2.9. Principal centres for coherent lidar measurements of atmospheric backscatter β(π, 10 μm) and a selection of more recent references.

Centre	Lidar Technique	References
NOAA Wave Propagation Laboratory (WPL), Boulder, Colorado	Groundbased, pulsed	[45–48]
NASA Marshall Space Flight Center (MSFC), Huntsville, Alabama	Groundbased, pulsed Airborne, CW	[34, 47–54]
USAF Geophysics Laboratory (GL), Hanscom AFB, Massachusetts	Groundbased, pulsed	[55–58]
Jet Propulsion Laboratory (JPL), Pasadena, California	Groundbased, pulsed Airborne, pulsed	[54, 59–63]
DERA (Malvern), UK—formerly RSRE	Airborne CW to 16 km altitude	[47, 48, 55–57, 64–67]
Several other smaller data bases, also SAGE extinction data		[39, 68, 69]

Research Agency, Malvern) conducted across the North and South Atlantic, and the NASA-supported GLOBE (Global Backscatter Experiment) programme, which extended across the Pacific. These studies were conducted well after the El Chichon volcanic eruption of 1982 and prior to the Pinatubo eruption of 1991. The data were thus accumulated in an historically 'clean' atmospheric period, and may thus represent the lowest levels of backscattering. This concept of a 'background mode' of atmospheric backscattering had in fact been suggested by Rothermel *et al* [47] from data drawn from two widely separated areas—over the UK and over the continental USA.

The measurements taken in the SABLE and GABLE trials with an airborne 10.6 μm cw lidar have been extensively analysed and discussed [56, 57] and the median and upper/lower quartiles and deciles are shown in figure C3.2.8. These appear to comprise the most comprehensive database currently available at 10.6 μm for the Atlantic region. The measurements amounted to nearly 200 000 individual records of backscatter made during 80 flights over the Atlantic in six different regions and/or seasons. They are put forward as a reasonable measure of global backscatter under generally clean atmospheric conditions at 10.6 μm.

At 1.54 and 2 μm, validated data sets of β(π) were collected over the Pacific Ocean during the GLOBE II mission in spring of 1990 [44, 70] and at 2 μm over the continental US in 1995–1996 [71], respectively. Measurements at 0.35–1 μm have been mainly directed to the estimation of the typical particle size in specific areas. However, measurements made with the NASA/LITE (Lidar In-space

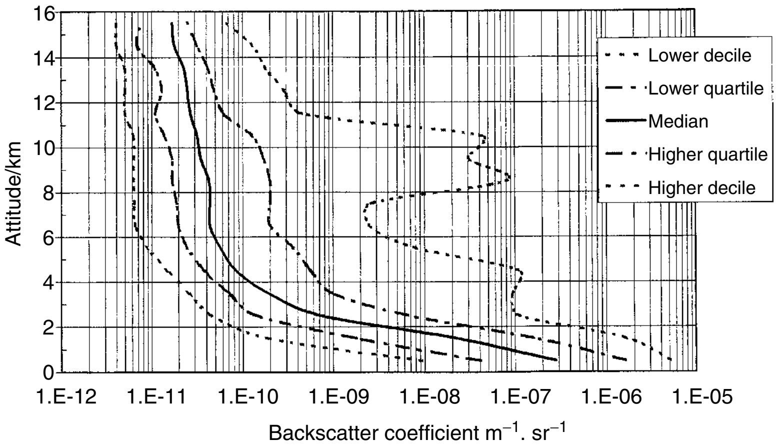


Figure C3.2.8. Median, quartiles and deciles of backscatter coefficient versus altitude at 10.6 μm as measured across the tropical mid-latitude and northern Atlantic and presented as a reasonable global measure under atmospherically clean conditions.

Technology Experiment; see also section C3.2.9.2) in September 1994 [72] have been able to document large areas of aerosol in the upper troposphere and stratosphere [73], and the transport of anthropogenically produced aerosols [36], Saharan dust [74] and biomass burning aerosols [32]. The measurements were made at three different wavelengths: 0.355, 0.532 and 1.064 μm. These three wavelengths were used to derive the aerosol and cloud characteristics.

Much effort has been put into the different comparison of backscattering at different wavelengths (see, e.g. Refs. [34, 51, 66, 67]). On the one hand, there are very few colocated and contemporaneous measurements of backscatter at different wavelengths. On the other hand, calculations based on Mie scattering are critically dependent on the assumed aerosol size distributions, chemical constitutions and refractive indices. Generally, the atmospheric aerosol backscatter coefficient increases with decreasing wavelength. For a wavelength range excluding anomalous refractive index changes, this dependence may be expressed by a power law:

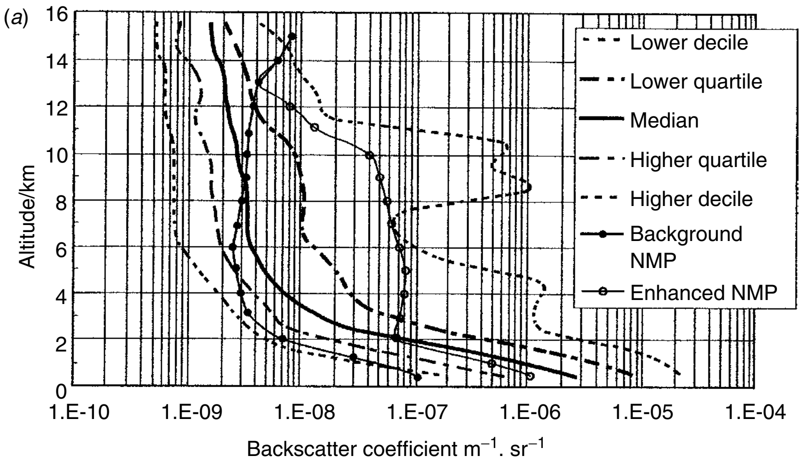
$$B(\lambda, z) = \beta(\lambda_0, z)(\lambda_0/\lambda)^{\varsigma} \tag{C3.2.21}$$

where $\beta(\lambda_0, z)$ is the backscatter coefficient at a chosen wavelength λ_0 and at a height z and where ς is the wavelength scaling exponent to derive backscatter at other wavelengths λ.

From the SABLE/GABLE data shown in figure C3.2.4, and comparison with other data including modelled Mie scattering [75] and other measurements at shorter wavelengths (e.g. Refs. [44, 54, 71]), the following expression for ς was derived:

$$\varsigma(0.35 < \lambda < 2.1\,\mu\text{m}, z) = 0.24|\log_{10}(\beta_0(10.6\,\mu\text{m}, z))| - 0.62 \tag{C3.2.22}$$

This expression is considered valid for shorter wavelengths, in particular the range 0.35–2.1 μm and may be used to translate the SABLE/GABLE data at $\lambda_0 \equiv 10.6$ μm. Just such a translation from 10.6 μm is shown in figure C3.2.9*a* and compares well with the measurements and other data at 1.06 μm shown in figure C3.2.9*b*. Another extensive and valuable investigation has been reported by Srivestava *et al* [76].



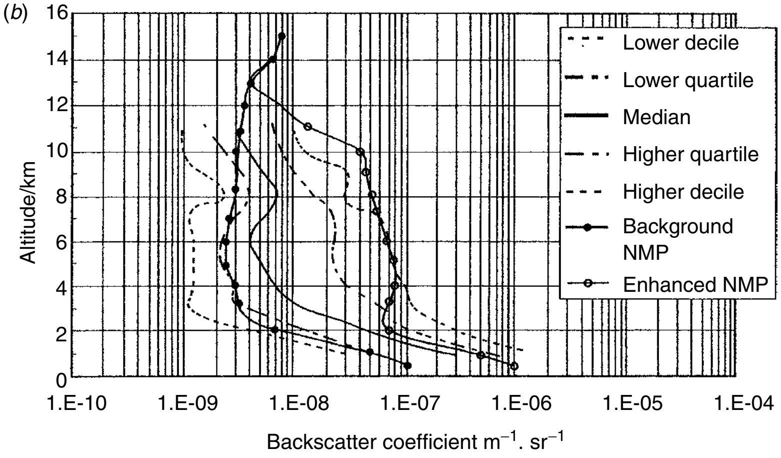


Figure C3.2.9. (*a*) Backscatter coefficient versus altitude at 1.06 μm as derived from figure C3.2.8 (for 10.6 μm) with the wavelength scaling of equation C3.2.21. The background and enhanced NASA New Millennium Programme (NMP) profiles are also included for comparison (see text). (*b*) Backscatter coefficient versus altitude at 1.06 μm for the GLOBE II measurements over the Pacific. The background and enhanced NMP profiles are also included for comparison.

Equation C3.2.22 must be used with caution. At 9.1 μm, for example, the situation is rendered particularly complex by the well-known 'resonance' in scattering due to the sharp changes in refractive index at this wavelength, particularly for ammonium sulphate; see, in particular, figure 3 of Srivestava *et al* [51]. In consequence, the backscatter at 9.1 μ is generally greater than at 10.6 μm and the ς component in the scaling law increases more sharply than at shorter wavelengths. The following equation encapsulates such an increase:

$$\varsigma(9.11\,\mu\mathrm{m}, z) = 1.25|\log_{10}(\beta_0(10.6\,\mu\mathrm{m}, z))| - 8.25 \qquad \text{(C3.2.23)}$$

A final example of aerosol backscatter measurement and comparable atmospheric trajectory analyses is shown in figure C3.2.10. Air masses with notably different aerosol backscatter, measured at three levels, 6.5, 8.25 and 10 km over the Arctic, originated over Europe, the United States and Canadian Arctic, respectively (from Ref. [77]).

(c) *Scattering from the Clouds*. Clouds display large backscatter coefficient (with $\beta_c(\pi) \approx 10^{-6}$ to $10^{-2}\,\mathrm{m}^{-1}\,\mathrm{sr}^{-1}$) and large extinction coefficient ($\alpha \approx 10^{-5}$ to $10^{-1}\,\mathrm{m}^{-1}\,\mathrm{sr}^{-1}$). The strength of scattering may vary by several orders of magnitude at a given wavelength and may vary by one to several orders of magnitude for different cloud size distributions and particle number density as a function of wavelength. Dense clouds strongly attenuate at laser wavelengths so lidar measurements cannot penetrate more than tens or hundreds of metres. However, holes are present in dense clouds at the small scale (on the order of 1–100 m) and results in so-called optical porosity. Such an optical porosity allows profiling of the atmosphere beyond dense clouds depending on the relative size of the lidar footprint and optical porosity (specifically on the probability density function of optical porosity). Lidar measurements at an angle close to nadir are favourable to reach the surface, although the probability of a clear line of sight through dense clouds does not decrease too much for angles not larger than 30–40° with respect to nadir. The only cloud type with an expectation of continuous lidar penetration is cirrus. The typical thickness of cirrus clouds is <1 km but on occasion thicknesses of up to 4 km have been observed.

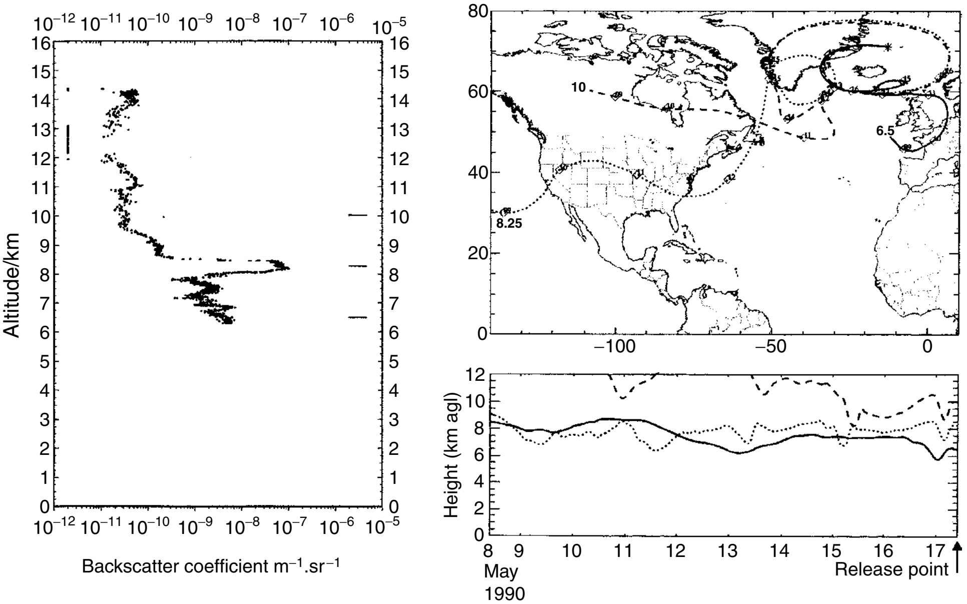


Figure C3.2.10. Flight 63 from Keflavik: altitude record of backscatter for the descent flying to the E along 71°N (in the region of 13°W) from 09.50 to 10.30 UTC on 17 May 1990. The strong scattering in the lower levels of the troposphere (<9 km) shows a very strong narrow band centred at 8.25 km which was absent on other legs of the flight. There was a sharp temperature inversion at 11.5 km; in the lower strasphere the scattering peaked at ~14 km but was very weak in the band 12–13 km. Air mass back trajectories for 9 days computed over the release point 71°N, 13°W at 10.15 UTC (May 1990) for 6.5, 8.25 and 10 km. Note the anticylonic loops and the remarkably different origins of the air masses for these trajectories from, respectively, Europe, the United States and the Arctic.

Table C3.2.10. Approximate values of extinction and backscatter coefficient of different cloud types in the ranges ~0.4–2 µm and ~10 µm (PSC—polar stratospheric cloud).

	Backscatter coefficient B_c (m^{-1} sr^{-1})		Extinction coefficient α_c (m^{-1})		
Cloud type	< 2 µm	Low 10 µm High	< 2 µm	Low 10 µm High	Altitude (km)
Cumulus	6.0×10^{-4}	1×10^{-5}–1×10^{-4}	1.2×10^{-2}	5×10^{-3}–30×10^{-3}	2–10
Stratus	5.0×10^{-3}	3×10^{-5}–5×10^{-4}	9×10^{-2}	1×10^{-4}–7×10^{-2}	0.2–0.7
Alto-stratus	1.0×10^{-3}	1×10^{-5}–1×10^{-4}	1.8×10^{-3}	3×10^{-3}–2×10^{-2}	2–4.5
Cumulo-nimbus	1.0×10^{-2}	4×10^{-5}–1×10^{-3}	1.8×10^{-1}	1.5×10^{-2}–6×10^{-2}	2–4
Cirrus	1.4×10^{-5}	1×10^{-6}–1×10^{-5}	2×10^{-4}	5×10^{-4}–5×10^{-3}	8–16
PSC	3.0×10^{-7}		6×10^{-6}		

An analysis of probability of occurrence of layered cirrus has been discussed by Vaughan *et al* [66] together with calculations of the impact on spaceborne lidar operation. Backscatter and extinction coefficients have been calculated for water clouds using Mie theory or relevant approximations for spherical particles with a good accuracy (e.g. Ref. [78]). This work has been conducted to derive the relevant relationships between microphysical parameters like liquid water content and optical parameters, i.e. $\beta(\pi)$, α, k (e.g. Refs. [79, 80]). As shown in table C3.2.10, backscatter coefficients for a given cloud type do not change sharply in the visible and near IR up to 1–2 µm, but are typically smaller by 1–2 orders of magnitude and more variable in the 10 µm region.

C3.2.4.3 Wind and related measurement

(a) Short range up to ~1 km

Wind measurement may be conducted with both pulsed and continuous wave (cw) lasers. The former are typically used for ranges > ~1 km and, with standard techniques of range gating, the range resolution is determined by the laser pulse length and signal processing. For a cw beam, the atmosphere of course provides an extended target with scattering at all distances. Nevertheless, as discussed in section C3.2.3.4, quite a sharp range resolution can be attained with a cw lidar by focusing the beam to give a peak sensitivity around the focal range F (figure C3.2.6).

An early measurement with a cw CO_2 LDV lidar is shown in figure C3.2.11. This record was made under rather gusty conditions on an airfield with the LDV pointed at an elevation angle of 30° into the prevailing wind; the Doppler spectrum was measured every 12.8 ms. The Doppler shifted frequency of the strongest signal in the spectrum was converted to an analogue voltage and displayed as a function of time. In figure C3.2.11*a*, the wind component changed from 7 to 17 m s^{-1} (from about 15 to 35 knots) over a period of about 5 s. Greater fluctuations are apparent at the shorter range in figure C3.2.11*b* due to greater wind turbulence at the lower height and the reduced spatial averaging and integration of the signal over the considerably shorter probe volume.

Such a system provides a single line-of-sight wind-velocity component. However, simple extension of such equipment enables one to determine the wind field anywhere around the measuring station. The laser beam may be scanned in order to resolve various components of the wind and several different scan patterns have been developed to suit particular measurement tasks. One of the simplest is to use a conical scan about a vertical axis. In this case, if the half-angle of the cone is φ (i.e. the angle of the lidar beam

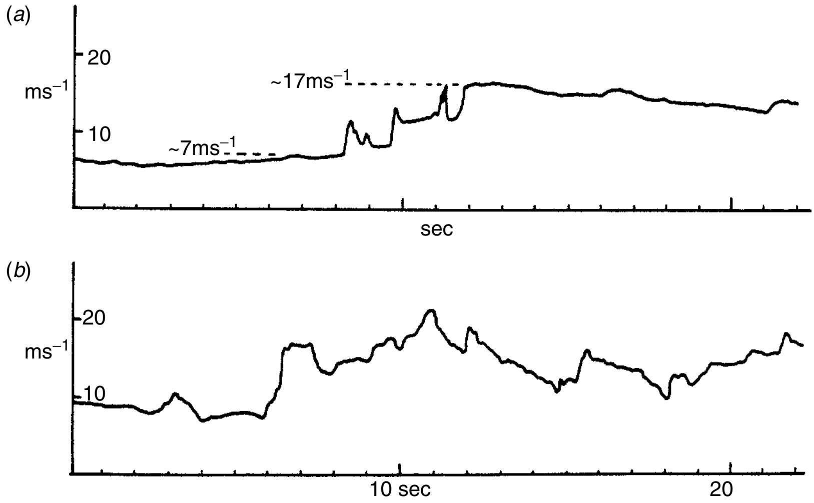


Figure C3.2.11. Records of strong gusts showing the line-of-sight wind component (derived from the dominant frequency in the Doppler spectrum) versus time. The CW CO_2 lidar beam was pointed at an elevation angle of 30° and the probe volume was set at a range of (*a*) 400 m and (*b*) 100 m.

from vertical) and θ_{sc} is the beam direction in the horizontal plane, the measured line-of-sight wind component V_M is given by:

$$V_M = V_H \cos(\theta_M - \theta_{sc}) \sin \Phi + V_v \cos \varphi \tag{C3.2.24}$$

where V_H and θ_H are the horizontal wind speed and bearing and V_V is the vertical wind speed. Examination of this expression shows that, as the beam rotates around the cone, any vertical up/down draughts (usually small) contribute a constant term $V_V \cos \Phi$ and the horizontal component varies with the cosine of $(\theta_H - \theta_{sc})$. This is readily demonstrated on a polar plot of V_M (or equivalent Doppler shift), as shown by the measurement example of figure C3.2.12. As the beam rotates, the magnitude of V_M traces out a figure-of-eight plot, for which the main axis gives the horizontal wind bearing (at θ_{sc} equal to θ_H) and the amplitude gives the wind speed. For the example shown in figure C3.2.9, the cw CO_2 lidar beam was focused at five successive heights from 25 to 250 m with three revolutions of the beam at each height. Each revolution was completed in 1 s with recording of 64 individual Doppler spectra. With less than 2 s to alter focus, the complete cycle of wind measurement from 25 to 250 m may be completed in less than 25 s. Rapid analysis of the Doppler spectra, and fitting to equation C3.2.24 to extract the wind parameters V_H, θ_H and V_V, may be carried out in real time. Figure C3.2.12 provides a clear example of changes of wind speed and direction with height and the data were acquired with a compact and highly mobile equipment mounted in a Land Rover vehicle, which could make measurements within 5 minutes of arrival at a site. Another example of measurement with a cw LDV is illustrated in figure C3.2.13. The conical scan was set to examine the changes of wind speed typically experienced by the revolving blades of a wind power generator. Comparison with simple models of turbulence showed reasonable agreement. Important questions of representativity and wind-velocity measurement errors for such systems have been considered experimentally and theoretically by Banakh *et al* [81].

Several ground-based cw systems have now been built worldwide, mostly employing CO_2 lasers with power output in the range 4–20 W. LDV systems based on shorter wavelengths, notably with cw

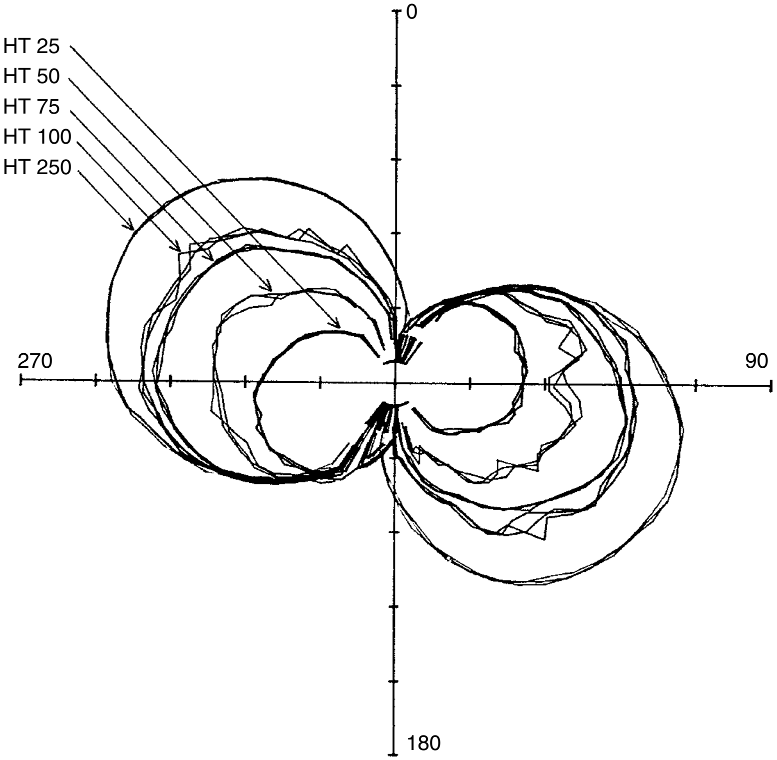


Figure C3.2.12. Polar plots of wind component at heights recorded at night in relatively calm conditions with a conically scanned CW CO_2 LDV. The wind strength (shown by the size of the figures-of-eight, with one scale division equal to $2\,\mathrm{ms}^{-1}$) increased steadily with height from $\sim 3.5\,\mathrm{ms}^{-1}$ (25 m) to $\sim 8\,\mathrm{ms}^{-1}$ (250 m). A strong directional wind shear (shown by the axes of the wind plots) is also very obvious and amounts to $\sim 36°$. Data recorded with the compace, mobile DERA (Malvern LDV).

diodes and fibre amplifiers in the 1.5 μm region, also offer much promise and are being developed for both monostatic and bistatic operation [82]. Such lidars should be compact, do not need detector cooling and incorporate fibre optic techniques for ease of assembly and precise wave-front matching. A recent example has been described by Karlsson *et al* [83] and used to investigate signal statistics of particulate scattering in the atmosphere [82]. The non-Gaussian character with high SNR from single particles is dramatically different from the complex Gaussian scattering with multiparticles in the probe volume (see also work by Jarzembski *et al* [84], who employed known particle sizes for lidar calibration).

The many applications of cw LDV include wind-flow measurement around buildings, smoke plumes from power stations and wind measurement from an oil rig to compare with satellite observations of the sea surface. Another potentially important application is towards wind-turbine power generation; Vaughan and Forrester [85] identified three specific local tasks in addition to the strategic benefits of a global wind measuring system (see section C3.2.9.3).

- *Site studies of local wind.* Wind generators require to be optimally sited. Conventional techniques with mechanical anemometers on tall towers can be very expensive. LDV, with its speed and mobility, could have a role.

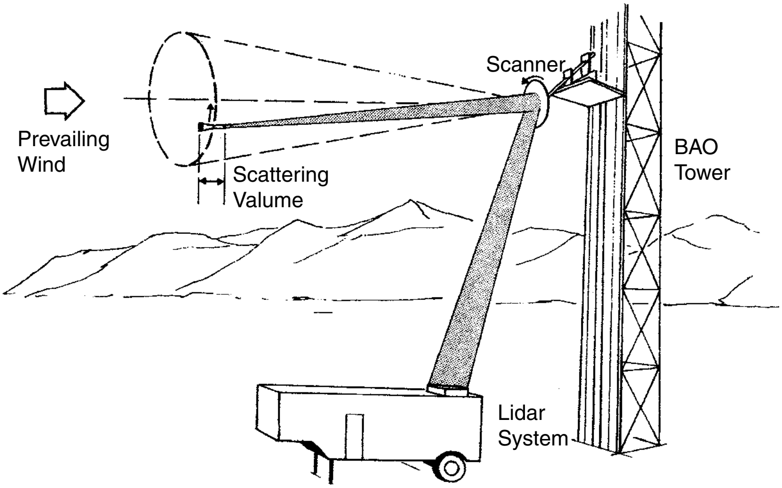


Figure C3.2.13. Schematic diagram for LDV measurements of rotationally sampled winds from a scanning mirror. The cone angle was fixed at 10° and the radius of the vertical measurement circle was varied by altering the beam focus (from Ref. [190]).

- *Complex flow around wind generators*. LDV offers a unique capability for rapid, non-intrusive measurement in quite small probe volumes within the complex flows generated by the turbine blades.
- *Wind monitoring and gust warning*. Measurements made by scanning up to hundreds of meters ahead (e.g. with LDVs mounted on the nacelle of the generator) would permit advance control of blade pitch for most efficient power extraction. In addition to routine gains in efficiency, such LDVs would also provide warning of occasional large gusts or turbulence; with appropriate blade control, the risk of consequent catastrophic damage to blades or gear-train would be reduced.

Other applications of cw LDV lidars for aviation (wake vortices, true airspeed, etc) are outlined in section C3.2.5 and military tasks in section C3.2.7.5.

(b) Longer range measurements

Laser Doppler wind measurements at ranges typically greater than ~1 km have been undertaken with several types of pulse lidar system:

- *Coherent heterodyne* at ~10 μm with CO_2 lasers, ~2.1 μm with Tm, Ho: YAG-YLF lasers, 1.55 μm with doped fibre amplifier lasers and 1.06 μm with Nd-YAG lasers. These all utilize aerosol scattering.
- *Interferometric direct detection* at 1.06, 0.53 and 0.35 μm with Nd-YAG lasers, utilizing both aerosol and molecular scattering.

For completeness, one should also note non-Doppler techniques, whereby time series of the strength of aerosol scattering from two beams of known separation may be compared (e.g. with correlation techniques). The time of passage of a strong fluctuation from one to the other (e.g. with peak in the correlation function) provides a measure of the cross wind but has limited temporal and spatial resolution [86–88].

Technology and application of *coherent LDV* have been well reviewed. For good frequency measurement, the transmitted laser beam requires to have minimal change of frequency through the duration of the pulse. For CO_2 lasers, the gas discharge conditions that ensure such good chirp characteristics, including gas catalysis, plasma effects and laser-induced media perturbation (LIMP), have been analysed by Willetts and Harris [89, 90]. Practical use of heterodyne detectors, and attainment of good performance, has been considered by many authors; in one study, near ideal behaviour with up to 13 dB local oscillators shot noise was demonstrated [91] (see also Ref. [92]). Signal processing with different algorithms and procedures has also been extensively reviewed (e.g. see Refs. [93–98]).

At the longer CO_2 wavelengths, TEA lasers have typically given larger pulses (0.2–3 J) with p.r.f. 0.1–10 Hz; for lower pulse energy (typically 1–10 mJ) and higher p.r.f, q switched and mini-MOPA lasers are employed. As an example at longer ranges and higher altitudes up to 10–20 km, the NOAA CO_2 lidar system has shown great reliability over many years. It has been used over widespread field campaigns with measurements of convective out flows, thunderstorm microbursts [99], sea breeze, flow in complex terrain [100] and downslope windstorms [101], etc. A record of measurement within the Grand Canyon taken at two levels below the rim is shown in figure C3.2.14. Differential flow along the bottom of the canyon established that, on occasion, pollutants entered from the southwest USA and contributed to wintertime haze [102]. At shorter ranges, and particularly for weak winds in the boundary layer, smaller pulse systems have been widely developed and used (see, e.g. Refs. [103, 104]).

A particularly long-ranging, coherent system at 1.06 μm was reported in 1993 by Hawley *et al* [105], employing a 1 J energy Nd-YAG coherent lidar. With aerosol scattering augmented by Pinatubo volcanic material, lidar measurements in a conical scan to 26 km altitude were documented, as shown in

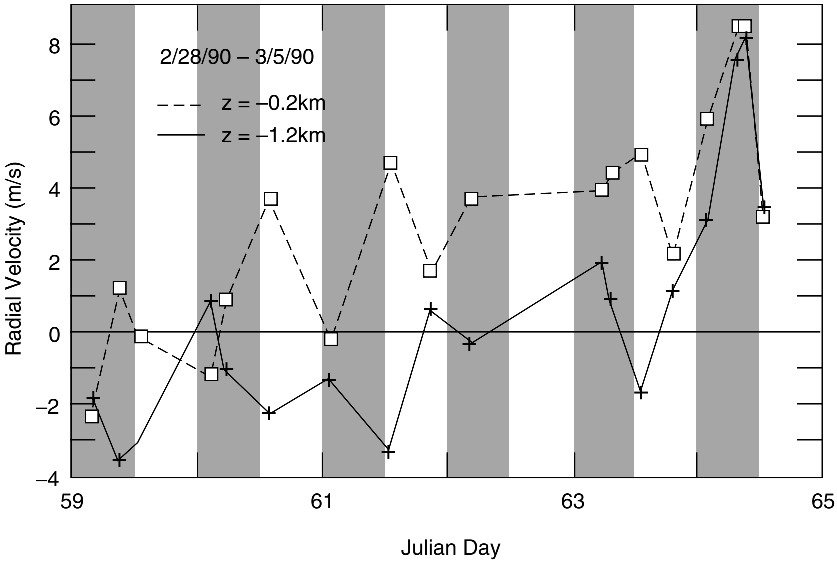


Figure C3.2.14. Radial flow within the Grand Canyon at two levels −0.2 and −1.2 km height relative to the canyon rim measured by CO_2 lidar. The shaded areas are night. Under the low wind (stagnant conditions) of the first few days, the flow at the top of the canyon is opposite in direction to the flow at the bottom of the canyon. In stronger winds at the end of the period, the flows were in the same direction.

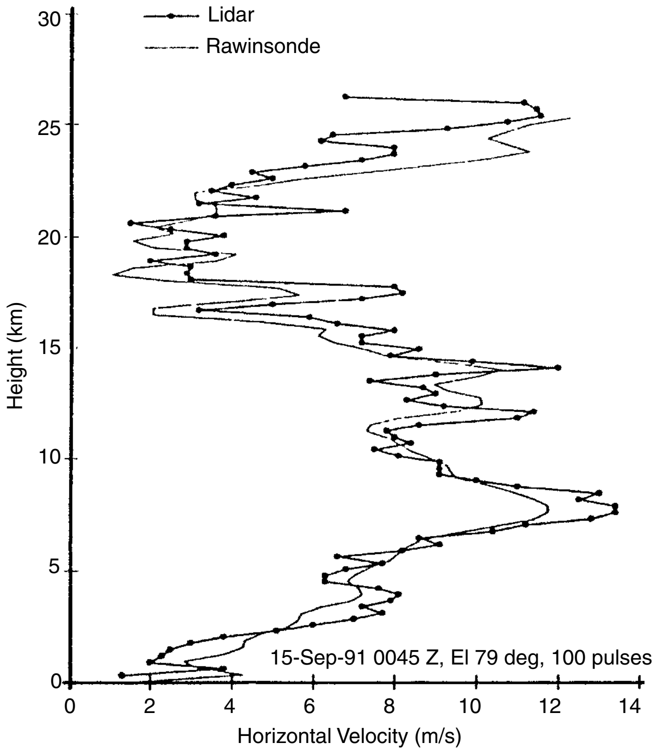


Figure C3.2.15. Comparison of wind speed with altitude for lidar and rawinsonde to 26 km measured above the Kennedy Space Centre Shuttle Landing Facility. The 1.06 μm lidar scanned at 20° off vertical at six fixed positions with 100 pulses averaging at each position. A complete record was completed in 3 min (from Ref. [105]).

figure C3.2.15, in good agreement with rawinsonde wind speed. Coherent lidar at the eye-safe 2 μm wavelengths has employed lasers with pulse energies, typically in the range 2–50 mJ. A plan view of a 2 μm coherent system is illustrated in figure C3.2.16. At the shorter wavelengths, range resolution may typically be 30–50 m (see section C3.2.3); measurements on a stationary hard target showed bias errors of $-3.3\,\mathrm{cm\,s^{-1}}$ with standard deviation error $11\,\mathrm{cm\,s^{-1}}$ [103]. Performance of a 2 μm coherent system was reviewed by Frehlich [106]. The compactness and potential for 'turnkey' operation are very favourable considerations and such systems are now commercially available. A pulsed lidar system based on fibre optic technology operating at 1–548 μ with an erbium doped fibre has recently been described by Pearson *et al* [107]. Operation in both bistatic and monostatic antenna configurations was investigated.

In addition to wind measurement with such lidars, it is worth noting that the strength of aerosol scattering may also provide useful information. One proposal has, for example, suggested the early detection of forest fires from the increased scanning due to quite minor smoke plumes. A lidar beam scattering from an elevated platform could provide coverage and monitoring over quite large areas.

As discussed in section C3.2.3, the design considerations for *direct detection, interferometric lidars* are very different. In particular, the signal may be accumulated from many small laser pulses without undue penalty (see figure C3.2.3) and across many optical modes. In addition, at shorter wavelengths,

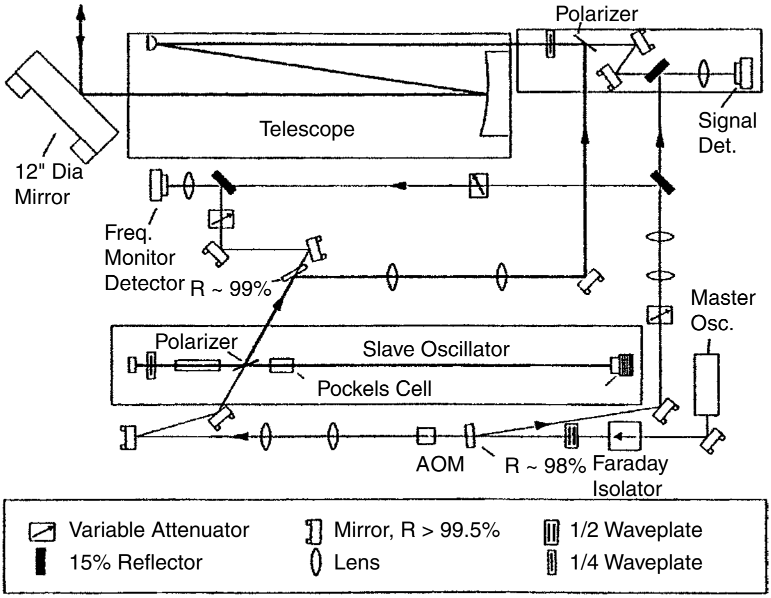

Polarizer
Signal Det.
Telescope
12" Dia Mirror
Freq. Monitor Detector
R ~ 99%
Master Osc.
Polarizer
Slave Oscillator
Pockels Cell
AOM
R ~ 98%
Faraday Isolator
Variable Attenuator
15% Reflector
Mirror, R > 99.5%
Lens
1/2 Waveplate
1/4 Waveplate

Figure C3.2.16. Plan view of a 2 μm injection-seeded coherent lidar reviewer (from Ref. [103]).

the molecular scattering is stronger and may be utilized. A number of different interferometric techniques have been developed:

(1) *Fringe imaging* (also called 'multi-channel') with multi-beam interference fringes for both aerosol and molecular scattering using either Fabry–Perot or Fizeau interferometers.

(2) *Double edge* (more strictly 'dual channel') with two pass-band Fabry–Perot filters set on either side of the scattered spectrum.

(3) *Visibility curve* with two-beam interference fringes formed in a Mach–Zehnder configuration.

Considerable controversy has been generated in the literature by the proponents of each technique. Theoretical calculations typically show calculated performance within a factor of 2–3 of the Cramer–Rao limit (discussed in section C3.2.3.2; see, e.g. Ref. [108]). Several fringe imaging and double-edge systems have been built and measurements made high into the atmosphere. The visibility curve (two-beam) technique is more recent but shows promising performance (see Ref. [109]).

In the fringe imaging technique, the spectrum of the scattered light is imaged across a number of detector elements (typically 10–20). A high-resolution system is obviously required for analysis of the narrow aerosol return (see figure C3.2.7) and the molecular spectrum is spread as a background noise across all the detector elements. The actual width of the aerosol signal is likely to be dominated by the laser line width itself. If this is of the order of 100 MHz (equivalent to ~20 m s^{-1} Doppler width at 355 nm wavelength) then the Cramer–Rao limit discussed in section C3.2.3 shows that, for a measurement accuracy of 1 m s^{-1}, the minimum number of photons required is given by

$$N_\mathrm{p} \geq (\eta T_\mathrm{t})^{-1}(0.425)^2(20)^2 \qquad \text{(C3.2.25)}$$

where ηT_t is an overall instrumental efficiency made up of η-detector efficiency and T_t instrumental transmission. Supposing $\eta T_t \approx 0.05$ at best gives a minimum input requirement at the receiving telescope of

$$N_p \geq 1.5 \times 10^3 \text{ photons} \tag{C3.2.26}$$

neglecting the impact of molecular background, daylight scattering, detector noise, etc. A lower resolution system for analysis of the much wider molecular Rayleigh line (of equivalent width ~600 m s^{-1} at 355 nm) with similar assumptions would require a minimum of ~1.3 × 10^6 photons.

Several fringe imaging lidars have been built and deployed in field trials (see, e.g. Refs. [110, 111]). Operation has usually been at 532 nm with frequency doubled Nd-YAG lasers. A particularly powerful system with 1.8 m diameter telescopes has been deployed at the ALOMAR Observatory (Andoya Rocket Range, Andenes, Northern Norway 69°N, 16°E). It has been used regularly for stratospheric wind measurements by day and night, through summer and winter (see Ref. [112]).

In the double-edge technique, Doppler shifts are derived by calibration of the response function R given by

$$R = (I_A - I_B)/(I_A + I_B) \tag{C3.2.27}$$

where I_A and I_B are the measured signals passing through the two fixed filters A and B placed on either side of the scattered light spectrum, as shown in figure C3.2.17. As the spectral line shifts in frequency, the relative magnitudes of I_A and I_B change. The fixed filters are formed by two Fabry–Perot cavities of slightly different spacing. With suitably chosen frequency separation of the filters, operation on the molecular spectrum can be made relatively insensitive to the narrow band, often highly variable, aerosol scattering at the centre. Precise calibration of R also needs to take account of the exact form of the broad molecular spectrum, which varies with temperature through the atmosphere. Several double-edge systems have been built [113–115]; one of the first, that at Observatoire Haute Provence (OHP) in southern France, operating at 0.53 μm, has made extensive measurements into the lower stratosphere over several years [116].

A number of field trials comparing different LDV lidars have now been conducted. The VALID trial at OHP in July 1999 [120] brought together double-edge (0.53 μm), fringe imaging (0.53 μm) and heterodyne (2.1 and 10.6 μm) lidars for comparison with a microwave wind-profiler radar and balloon radiosondes. Forty-six data sets were accumulated and generally showed good agreement with cross-correlation coefficients above 60%. For some specific meteorological cases (e.g. a jet stream), measurement discrepancies were readily explained. At a trial in September 2000 at Bartlett, New Hampshire, USA, a fringe imaging lidar (0.53 μm), a double edge lidar (aerosol channel at 1.06 μm, molecular at 0.35 μm) and heterodyne lidar (1.06 μm) were deployed with microwave wind profiler and regular radiosonde launches. Generally reasonable agreement in a variety of atmospheric conditions was obtained.

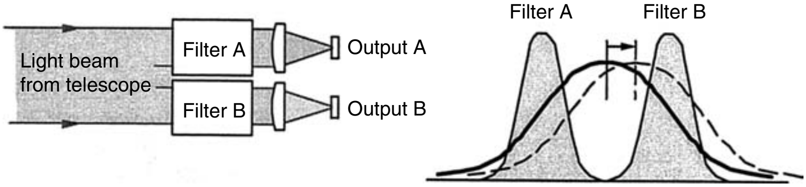


Figure C3.2.17. Schematic of the double edge technique with two filters A and B set on either side of the broad molecular spectrum as utilised for scattering at 0.35 and 0.53 μm. The technique can also be applied to the relatively stronger aerosol channel at 1.06 μm with appropriately spaced filters at higher resolution.

Table C3.2.11. Application of cloud measurement data (height, structure etc).

Climatological studies	Radiation balance Global energy balance Aqueous phase chemistry Monitoring of polar stratospheric clouds and ozone studies
Weather forecasting	Delineation of air mass discontinuities Assessment of frontal or convective activity Assimilation into numerical weather forecasting models, mesoscale, regional and global Improved quality of humidity and temperature retrievals Improvements in cloud vector winds and sea surface temperature
Local information	Objective assessment of cloud base and cloud cover, particularly for aviation and airport operation Precipitation and vertical visibility

C3.2.4.4 Cloud measurement

Detailed knowledge of cloud parameters, as indicated in table C3.2.11, is important for three major applications: climatological studies, weather forecasting and local information, particularly in aviation. In addition, knowledge of polar stratospheric clouds (PSCs) and their role in atmospheric chemodynamics, particularly for ozone destruction, is of great significance (see section C3.2.6). Laser radar can be used to measure cloud height, vertical extent, structure, optical thickness, statistical distribution, classification, presence of thin and subvisual cirrus, precipitation, etc, and many studies have been conducted (see, e.g. Refs. [59, 118–121]). Spaceborne lidar systems obviously have the potential to provide such information on a global scale, as illustrated, for example, in the discussion following the first such demonstration of the NASA LITE equipment [72].

For local measurements, ground-based systems have been extensively developed and are commercially available. Relatively simple techniques with pulsed lasers are employed with direct detection of the scattered light along the beam in ranging intervals of typically 100 ns (15 m equivalent). In fair weather and well-marked clouds, the basic cloud height measurement is a straightforward ranging task. Consideration does need to be given to the precise definition of 'cloud height'—should it be the initial return from the lowest fluffs of cloud, the height of the steepest increase in backscatter or the height of maximum backscattered signal? In a typical commercial lidar ceilometer (e.g. the Vailsala CT25K), the processing algorithm selects a height in the region of peak backscatter—which corresponds to the height from which pilots can usually see the ground well.

In severe weather, the problem becomes much more difficult; stray signals will be evident from rain, snow, virga (precipitation out of cloud evaporating before it reaches the ground), haze and fog. Quite complex algorithms have been developed to invert the measured backscatter profile to provide a credible vertical extinction coefficient profile. Thresholding criteria based on experience and extensive observations can then be applied to the latter to determine and classify a cloud presence. As mentioned, the Vaisala CT25K lidar ceilometer is a robust, fully automated system based on a single lens, monostatic transmitter/receiver and an InGaAs pulsed diode laser operating at 905 nm. Measurements can be made from 0 to 25 000 ft (7.5 km) with 500 range gates and range resolution of 50 ft (15 m) in a programmable measurement cycle of 15–120 s.

This equipment has been further developed with partial funding by the European Space Agency and uses an array of four standard equipments operated in synchrony. Above 1000 m, all four beams overlap so that every receiver collects scattered light from every transmitted beam. The increased SNR ensures that this more powerful equipment provides a measurement capability on cloud up to the 75 000 ft (21 km) altitude referenced by the World Meteorological Organisation (WMO) in 1983 as the upper limit for cloud height observations.

C3.2.5 Lidar and aviation

Following the demonstration of lidar measurements in the atmosphere, the application of such techniques to problems in aviation was rapidly developed. Immediately obvious topics were measurement of true airspeed, warning of wind shear and turbulence, calibration of pitot-static pressure-differential probes and the potential hazards of aircraft wake vortices. For operation in aircraft, equipments must obviously be compact, reliable and robust to the environment of vibration and reduced pressure.

C3.2.5.1 True airspeed, wind shear, turbulence and pressure error calibration

In the late 1970s, a number of airborne Doppler lidars were built. One of the earliest was quite a large-pulsed CO_2 system designed to look forward of the aircraft at ranges > 5 km. One of the main aims, to establish whether the equipment could detect clear air turbulence at high levels in the atmosphere, was successfully demonstrated (see, e.g. Refs. [49, 103]). For measurements at shorter ranges, the Laser True Airspeed System (LATAS) was built at the Royal Signals and Radar Establishment (RSRE) in 1980 and flown in aircraft of the Royal Aircraft Establishment (RAE). Figure C3.2.18 shows the installation of

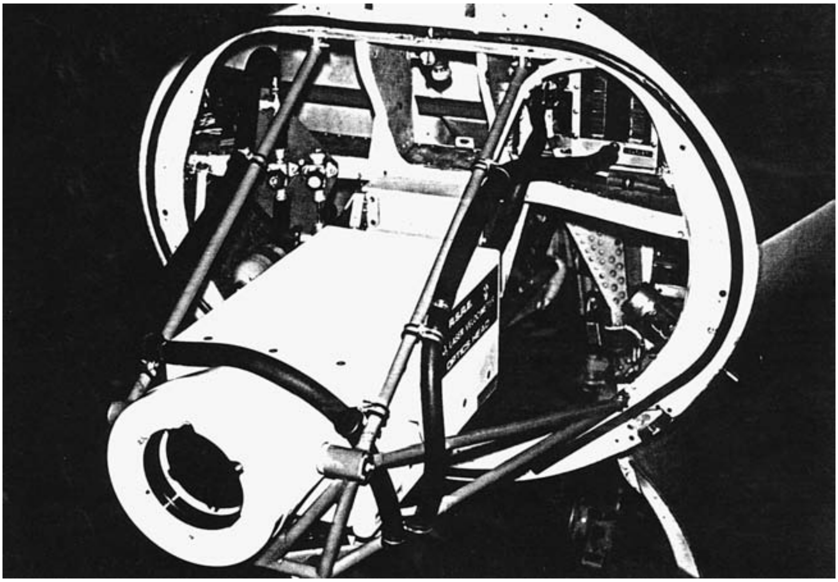


Figure C3.2.18. The Laser True Airspeed System (LATAS) installed in the unpressurized nose of an HS125 aircraft. For flight, this was covered with a nose cone holding a germanium window through which the lidar beam emerged.

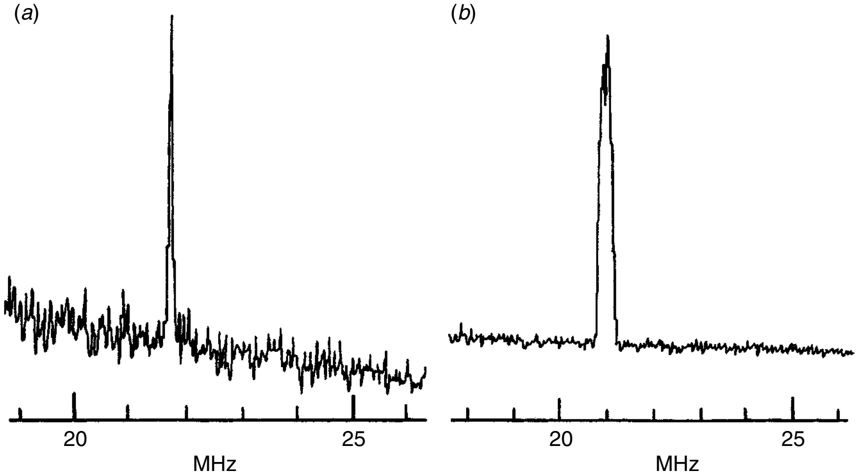


Figure C3.2.19. Typical Doppler spectra recorded in flight with the LATAS equipment at focal range of 100 m with 25 ms experiment time: (*a*) height 5000 ft, the peak frequency corresponds to 225.2 knots; (*b*) height 960 ft, the mean frequency corresponds to 215.0 knots; note the broadened spectral width due to the turbulence of ~5 knots.

LATAS in the unpressurized nose of an HS125 executive jet-type aircraft. The continuous wave CO_2 laser gave a power output of ~3 W and the optics head was contained in a well-insulated temperature-controlled enclosure. All controls were operated remotely from a console in the aircraft cabin. Figure C3.2.19 shows a pair of typical spectra recorded at different levels in the atmosphere; one shows increased levels of turbulence in the probe volume.

From the peak of such spectra, the true airspeed of the aircraft (that is, the speed of the aircraft relative to the air it is moving in) can be determined with an absolute accuracy of typically better than $0.2\,\mathrm{m\,s^{-1}}$. Figure C3.2.20 shows a record of such measurements made at $25\,\mathrm{s^{-1}}$ in the atmospheric

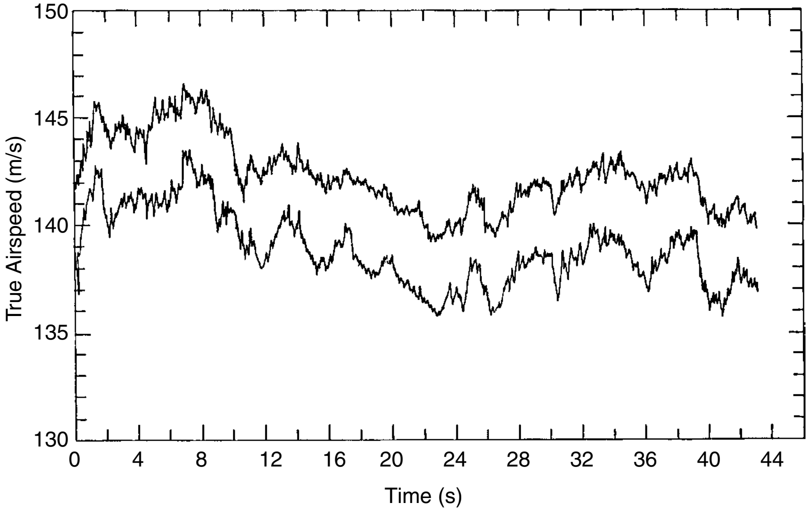


Figure C3.2.20. True airspeed recorded against time. The top trace is the lidar data with sampling at $25\,\mathrm{s^{-1}}$. The lower trace is the speed calculated from a pressure-differential gust probe mounted on a boom. Note the close correspondence of velocity structure (from Ref. [122]).

boundary layer [122]. For these studies, the lidar was focused at a reduced range of 17 m (but sufficiently far enough to be away from any disturbance of the advancing aircraft) to give good comparison with a small, conventional, pressure-differential gust probe mounted in a thin boom projecting ~1 m ahead of the aircraft. The air speed derived from the pressure-differential system is also shown in figure C3.2.20. The small difference between the two airspeeds (about 3 m s^{-1}) almost certainly arises from inaccurate correction of the pressure data, and provides the basis for calibrating such devices as discussed later. Closer examination of the data in figure C3.2.20 shows a time offset of ~0.11 s (clearly corresponding to the 16 m forward look of the lidar), and excellent agreement of the detailed structure (and computed variances) for the two sets of data.

Although clear air turbulence at high level in the atmosphere is troubling to aircraft, with risk of injury to unbelted passengers and crew, few if any aircraft have been caused to crash. Low-level wind shear, on the other hand, often but not always associated with thunderstorm activity, has caused many serious accidents, notably with aircraft in landing or take-off phase. At this stage, aircraft have a low airspeed, typically ~120 knots (~60 m s^{-1}). In the presence of wind shear, as the aircraft passes into a region of notably different wind speed, its lift may be greatly reduced. The pilot may have insufficient time to accelerate the aircraft up to sufficient airspeed to keep it aloft. Figure C3.2.21 shows an early LATAS record of passage through a thunderstorm microburst during the Joint Airport Weather Studies (JAWS) trial in Colorado in 1982. In this case, the focal range was set 250–300 m ahead of the aircraft. The sensitivity extended out to 700–800 m and thus strong shear or turbulent structures entering the extended probe at longer range were evident. In figure C3.2.21, the sequence of Doppler spectra in the lidar record showed a headwind that changed by over 40 kt (~ 20 m s^{-1}) in about 5 s. There was an additional down draft of ~6 m s^{-1}. Analysis of these measurements contributed to the development of a descending vortex-ring model for thunderstorm microburst behaviour, in contrast to the more usual vertical jet and outflow model. Simulations with the LATAS parameters have shown that the ~5–10 s warning of shear from a probe range of ~300–600 m could be useful if heeded promptly [123–125]. Wind shear and microbursts are dynamic phenomena; indeed, the simulations showed that there was significant advantage in controlling the aircraft (a medium-size passenger jet) using the airspeed measured ~300 m ahead, but increasing the distance to 600 m produced little further improvement. Nevertheless, in terms of general operation, it would be useful to look at a greater range with pulsed systems and a number of airborne equipments were built in NASA-supported programmes. Both CO_2 10 μm and solid state 1 and 2 μm lidars were successfully demonstrated at ranges to ~3 km. Comparative performance in terms of range and velocity resolution, and atmospheric factors in different conditions of rain, high humidity, etc, were assessed [126].

While wind shear remains a severe problem, the widespread deployment of lidars in civil airliners nevertheless appears unlikely in the near future. Increased awareness of the meteorological factors that precipitate severe wind shear, combined with the warning from large powerful Terminal Doppler Weather Radars (TDWR) operating at major airports, provide improved confidence and safety in airport operation. Nevertheless, it may be remarked that lidar has a very strong capability to detect so-called ‘dry’ microbursts (i.e. without associated rain) and is thus complementary to the high sensitivity of TDWR for conditions of rain and high humidity.

In an extensive development in France in the late 1980s, a CO_2 10 μm cw lidar was similarly used for true airspeed measurement and flown in a helicopter and various transport and fighter aircraft. This was subsequently configured into a three-axis equipment called ALEV-3 by Sextant Avionique, as shown in figure C3.2.22. This lidar has been used for air data calibration and aircraft certification in extensive flight tests of Airbus aircraft. With the beams focused at 70 m, the equipment gives a precise, real-time measurement of the airspeed vector and thus permits the calibration of static pressure and the angles of incidence and sideslip [127].

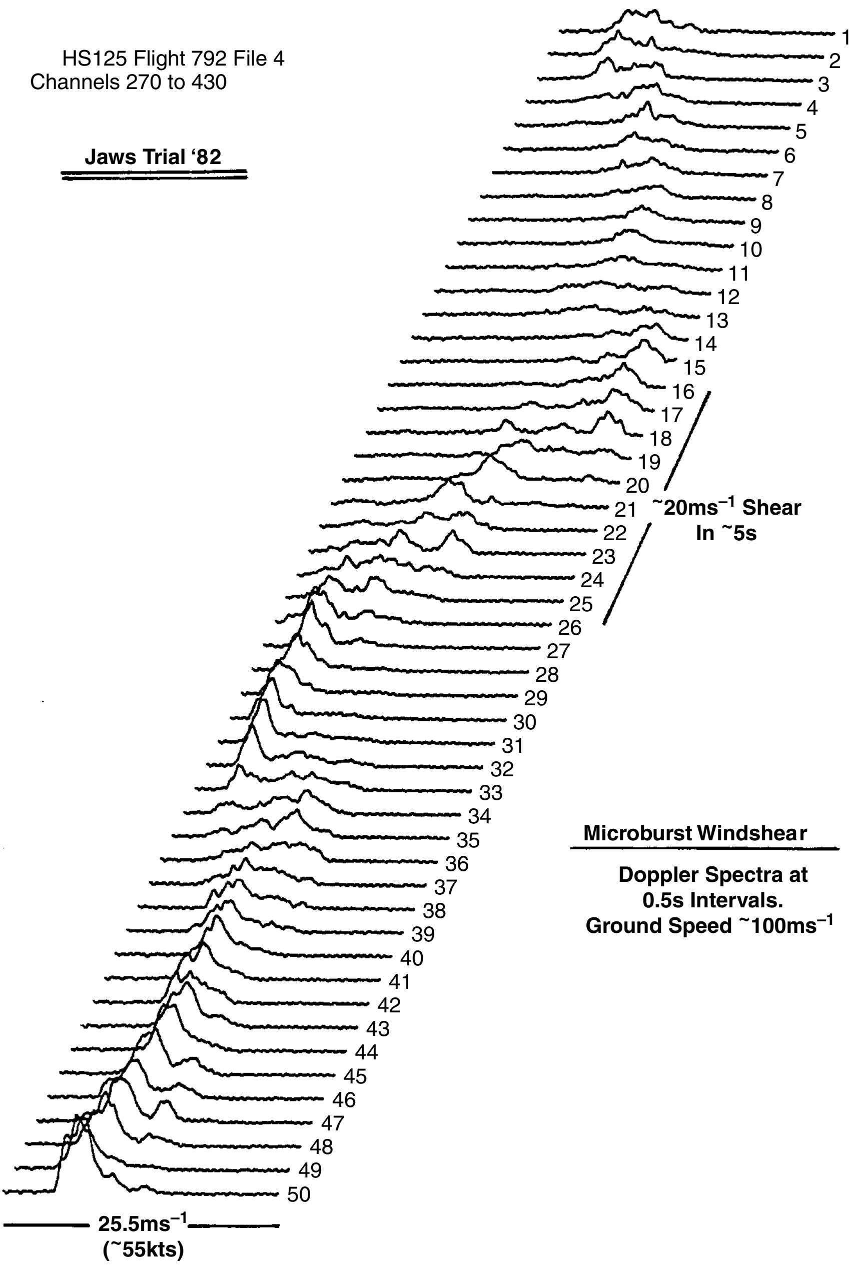


Figure C3.2.21. Sequence of lidar spectra recorded at 0.5 s interval during aircraft passage through a thunderstorm microburst. Note the severe wind shear (40 knots, ~20 ms^{-1} in ~5 s) in the record.

An advanced optical air data system (OADS) for measurements high in the atmosphere was built and flown in the early 1990s [128, 129]. Good performance at ranges to a few tens of metres was obtained with a direct detection technique in which the velocity components were derived from the measured transit times of aerosols between light sheets. This programme is outlined in Chapter C5 of the present volume and used as an exemplar of optical engineering practice in a hostile environment.

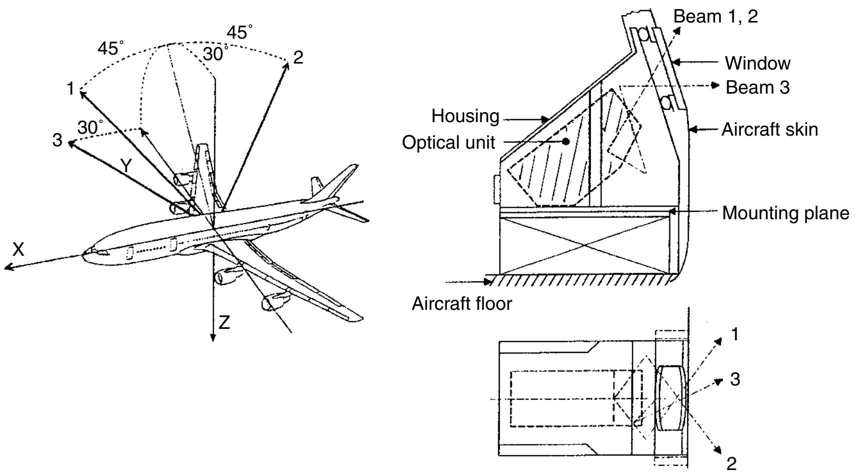


Figure C3.2.22. The ALEV-3 airborne lidar for air data calibration and aircraft certification. The optical axes chosen for transport aircraft. The lidar (optical unit) and installation in the aircraft (from Ref. [127]).

More recently, an airborne wind infrared Doppler lidar (WIND) has been developed in cooperation between CNRS and CNES in France and DLR-Munich in Germany [130]. In this lidar, built more for meteorological investigations and as a precursor for study of spaceborne systems, the TE-CO_2 pulse laser operates at ~300 mJ pulse and 4 or 10 Hz prf. Successful measurements from an altitude of 10 km, with downward conical scanning at 30° from nadir, have been undertaken in international field campaigns since 1999 [131].

C3.2.5.2 Aircraft wake vortices

In the process of generating lift (deriving from the pressure difference between the upper and lower surfaces of the wing aerofoil), all aircraft (including helicopters) create transverse rotational flow in the air that has passed over each wing. This rotational flow rapidly evolves into two powerful counter-rotating vortices that extend as a pair of trailing ribbons behind the aircraft. The initial separation of the vortices is about 80% of the wing span; their rotational sense produces a strong downdraft in the region between them, and in free air they tend to sink at a rate of 1–2 m s^{-1}. Their trajectory is, of course, largely determined by meteorological conditions; on approach to the ground, the sink rate is reduced and the vortices tend to separate. An extensive assessment and bibliography of the then state of knowledge of wake vortices was given by Hallock in 1991 [132].

The existence of such wake vortices, sometimes rather dramatically referred to as ‘horizontal tornados’, represents a potential hazard to following aircraft, particularly for smaller aircraft following larger in the vicinity of airports on landing and take off. The Air Traffic Control System guards against this hazard by the application of separation minima between all pairs of aircraft operating under Instrument Flight Rules. The larger and heavier the aircraft, the more powerful the vortex it creates and hence the separation distances depends on the types of aircraft (set into various weight categories) involved. Such separation minima provide a significant constraint on airport capacity—it has been estimated, for example, that if they were not necessary the arrival capacity at Heathrow could be increased by up to five aircraft per hour. Obviously, the separations must be sufficiently large to ensure safety, but they should not be so excessively conservative as to unduly restrict the capacity of airports.

The characteristics of wake vortices—their formation, evolution, persistence, trajectory, mode of decay, etc—have been studied for many years. In particular, powerful techniques of computational fluid dynamics and large wind tunnel measurements on reduced scale model aircraft provide information on the early stages of vortex formation and evolution. Recent water tank studies have extended this to the late stages. The positive and complementary features of lidar measurements are that they can be carried out on full-scale aircraft, in the real atmosphere and potentially at long distance (many 100 s of wing span) downstream of the generating aircraft and thus on fully mature and decaying vortices.

Wake vortices were, in fact, amongst the very first subjects of study by coherent laser radar. A demonstration of coherent 10 μm LDV measurements with cw lidars was made in 1970 in a NASA-sponsored programme. Subsequently, during the 1970s and early 1980s, very extensive NASA- and Federal Aviation Agency (FAA)-funded investigations were made and helped to define vortex separation standards [132]. In the 1980s, the German DLR Institute of Optoelectronic deployed a high-performance cw lidar at Frankfurt Airport between two parallel runways with the prime aim of examining the risk due to a cross wind possibly transporting vortices from one runway to the other [133]. The runway separation of 520 m is often too small for operating both runways independently with respect to wake vortices. Figure C3.2.23 shows a section of the vertical measurement plane with

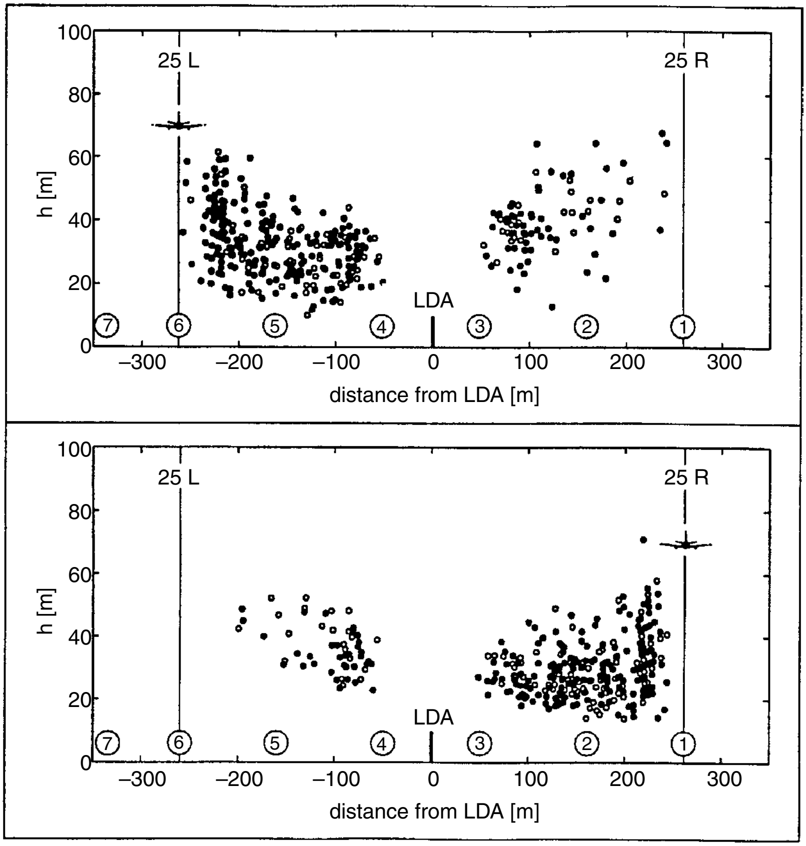


Figure C3.2.23. Position of wake vortices at Frankfurt airport measured by the Laser Doppler Anemometer (LDA) situated between the two parallel runways 25L (left) and 25R (right). The upper record shows the propagation of Boeing 747 vortices generated on runway 25L moving towards 25R, under the prevailing cross wind, and the lower from 25R in the opposite wind flow.

examples of vortices moving under the prevailing cross wind from one runway to the other. A number of such vortices show a steep ascent towards the parallel runway. This bouncing effect may enhance the hazard since the vortex will tend to pass near the altitude of approaching aircraft. This lidar work has contributed to the development of a strategy of operation for Frankfurt Airport in various cross wind conditions.

With increasing demand in the 1990s for available runways, it is clear that wake vortices and the required separation minima limit the capacity for the world's busier airports. There has thus been increased interest recently in lidar measurements with NASA and FAA programmes in the USA, Civil Aviation Authority-funded programmes in the UK and EC-funded programmes in Europe. This includes 10 μm cw lidars with controlled tracking of vortex position and comparison with acoustic sensors and mechanical anemometers. In addition, short wavelength 1 and 2 μm pulse lidars have been used for longer ranges.

A schematic illustration of the flow field and velocity distribution for vortices is shown in figure C3.2.24. The Doppler lidar spectra expected for a beam intersecting such a flow is somewhat complex due to the changing line-of-sight velocity component along the beam and the spatial weighting function of the lidar. The form generally expected is shown figure C3.2.5, as calculated by Constant *et al* [134]. Most importantly, scattering from the immediate tangent region provides the highest frequency peak component in the spectra. The intensity of these peaks rapidly weakens as the beam gets closer to the vortex core. Nevertheless, with a lidar of adequate spatial and spectral resolution, their measurement provides a velocity profile through the vortices, as shown in the series of experimental spectra in figure C3.2.25 with a fixed pointing lidar. In these pictures, the spectra are plotted vertically and coded for intensity as shown in the inset. The rapidly rising peak Doppler frequency on approach to the vortex core is very obvious.

With the lidar beam scanning to the side, many successive intersections on vortices with the characteristic cusp-like signatures may be obtained. Analysis of a record from Heathrow showed a very remarkable result in that the near-wing vortex from a B-747 aircraft pursued the unusual path shown in figure C3.2.26. After initially sinking to about 40 m above ground level, it subsequently rose and, at ~70 s, was back at the glide slope with essentially undiminished strength. This is a good example where the ATC separation standards worked well—the following aircraft passed through at ~102 s, well after the vortex had moved away. The unusual character of this example must be emphasized; in the overwhelming majority of the nearly 3000 lidar records at Heathrow, the vortices were rapidly convected away from the glide slope by the prevailing wind within 30–40 s.

Much quantitative data on vortex character, strength and longevity may be obtained from such lidar records. The strength of vortex circulation or vorticity $\Gamma(r)$ in $\mathrm{m^2\,s^{-1}}$ is given by the expression

$$\Gamma(r) = 2\pi r V(r) \tag{C3.2.28}$$

where $V(r)$ is the rotational speed at radius r. From figure C3.2.25, values of vorticity for the B757 vortex of $\Gamma(14.3\,\mathrm{m}) = 331\,\mathrm{m^2\,s^{-1}}$ (at mean age 16 s) and for the B747 of $\Gamma(22.4\,\mathrm{m}) = 545\,\mathrm{m^2\,s^{-1}}$ (at mean age of 17 s) are obtained, in reasonable agreement with expectation for these aircraft in landing configuration. As a further example, from the scanned record for figure C3.2.26, much detailed information may be extracted, including the values and changes of vorticity at different radii over the full observation period. The circulation $\Gamma(13.5\,\mathrm{m})$ remained close to $510\,\mathrm{m^2\,s^{-1}}$ out to 70 s [135].

In recent years, study of aircraft wake vortices has further intensified; EC-supported programmes have been very fruitful in combining lidar studies with wind tunnel, water tank, catapult launching of models and computer fluid dynamics (CFD) investigations (see, e.g. Ref. [136]). Pulse laser systems at 2 μm wave length and spatial resolution of ~50 m may also be used for localization and measurement of vortices at ranges greater than ~1 km, as first demonstrated by Hannon and Thomson [137]. Lidar

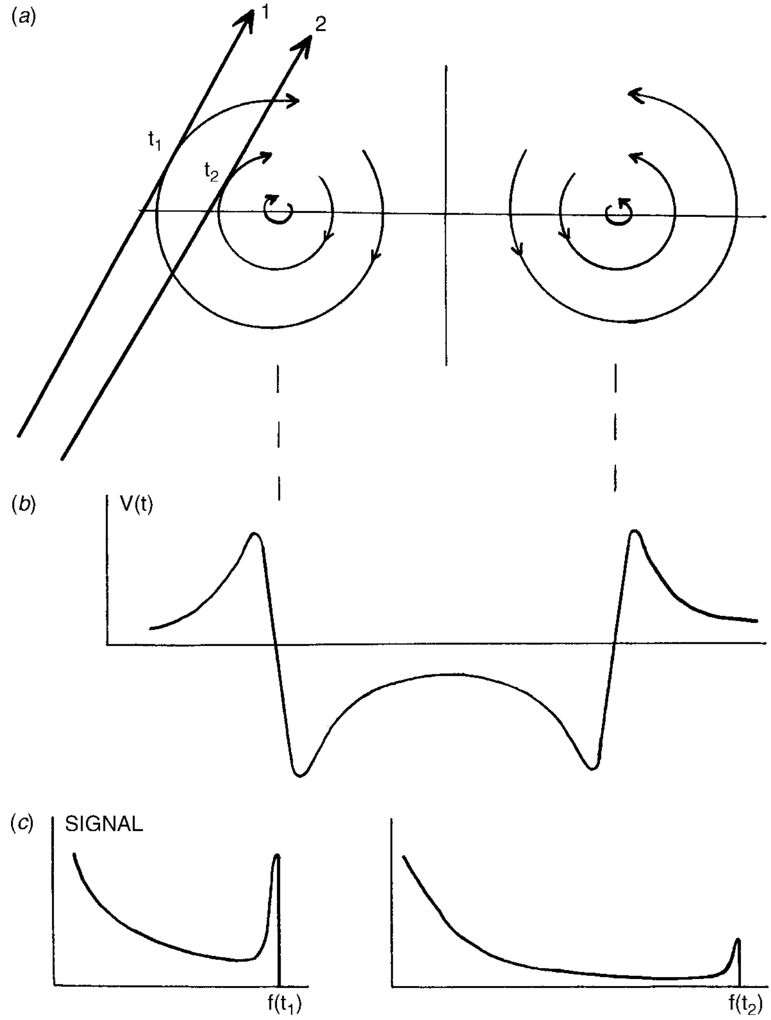


Figure C3.2.24. Schematic of vortex flow and lidar spectra. (*a*) Cross section of a pair of rotating vortices. Two positions A and B of a lidar beam intersecting the left hand vortex at tangent positions t_1 and t_2 are shown. (*b*) Approximate peak velocity profile $V(r)$ along the line between the vortex cores. (*c*) Doppler spectra for the lidar beams at A and B. Analysis shows that scattering from the immediate tangent regions t_1 and t_2 give the peaks in spectra at Doppler frequency $f(t_1)$ and $f(t_2)$ equal to $2V(t_1)/\lambda$ and to $2V(t_2)/\lambda$, respectively.

studies include measurement of military-type aircraft in Germany [138], report of measurements at Dallas/Fort Worth airport [139] and at Heathrow [140, 141]. Very recent collaborative campaigns with several lidar systems, both CW and pulsed, have been held at European airfields. Precise comparisons of wind tunnel data with lidar measurements have been made [142], and observations reported of an unusual structure with a vortex within a vortex [143]. A recent paper considered the impact on helicopter handling of vortex encounters using lidar data [144].

In summary, a broad view of the application of lidar to problems of aircraft wake vortices suggest several principal areas. First, at the most basic level, there is the important contribution to fundamental

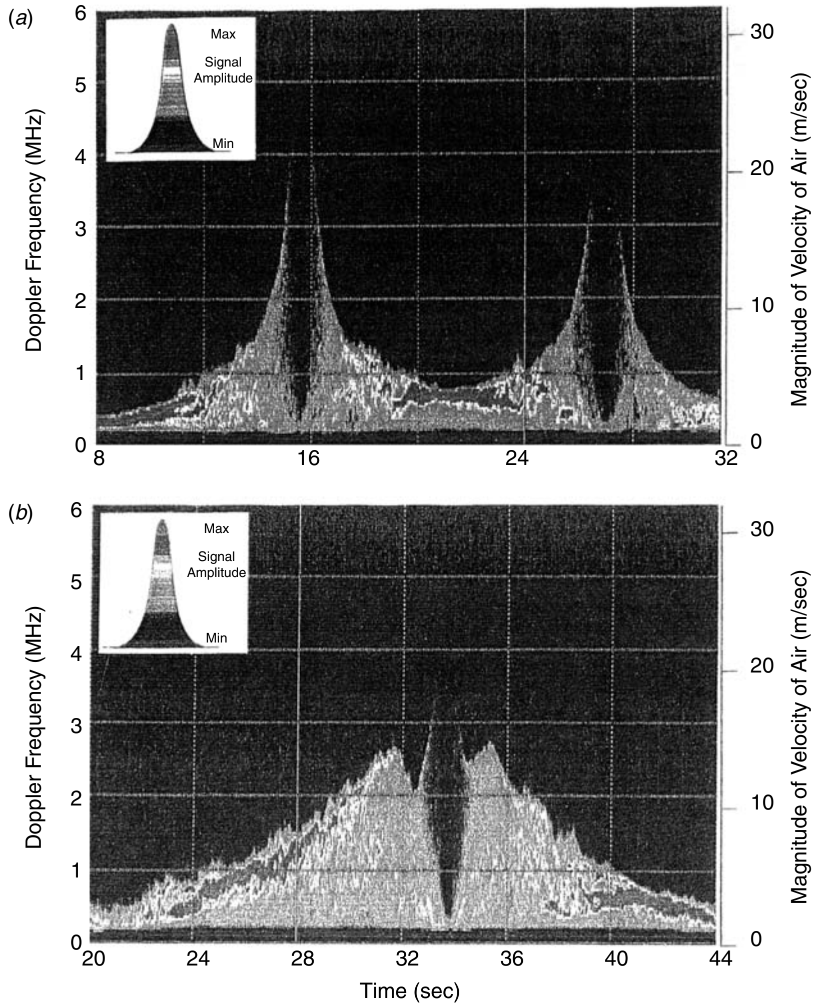


Figure C3.2.25. Doppler spectra (recorded at $25\,\mathrm{s}^{-1}$) from wake vortices carried by the prevailing cross wind through a static, upward looking lidar. Each spectrum is plotted vertically and colour coded for amplitude as shown by the insert. The time after passage of the generating aircraft is shown on the horizontal scale; the Doppler frequency (left hand) and equivalent velocity component (right hand) are on the vertical scales. (*a*) Pair of vortices from a B757 aircraft (CAA 320133, 3 October 1995); lidar elevation 90° and focal range 110 m. The cross wind was $4.9\,\mathrm{ms}^{-1}$; note the very sharp profiles of the vortices, rising close to the top of the scale, on either side of the cores. The downflow midway between the vortices is $7.4\,\mathrm{ms}^{-1}$. (*b*) Pair of vortices from a B747-200 aircraft (CAA 320068, 30 October 1995); lidar elevation 90° and focal range 115 m. The cross wind was $6.0\,\mathrm{ms}^{-1}$ and downdraft $\sim 0.7\,\mathrm{ms}^{-1}$. Note the characteristic shape with broad, steadily rising profile up to $\sim 12\,\mathrm{ms}^{-1}$ and then rounding over into a series of inversions. The corrected downflow V_v between the vortices is $7.78\,\mathrm{ms}^{-1}$.

knowledge of vortices—their formation, interaction, transport, decay and dissipation and comparison with data using other techniques. This potentially should input into detailed aircraft design (e.g. wing configuration, interaction of engine, flap and tip vortices, etc) with development of low vorticity vortices (lvv) or quickly dissipating vortices (qdv). Such progress is important particularly for the very large

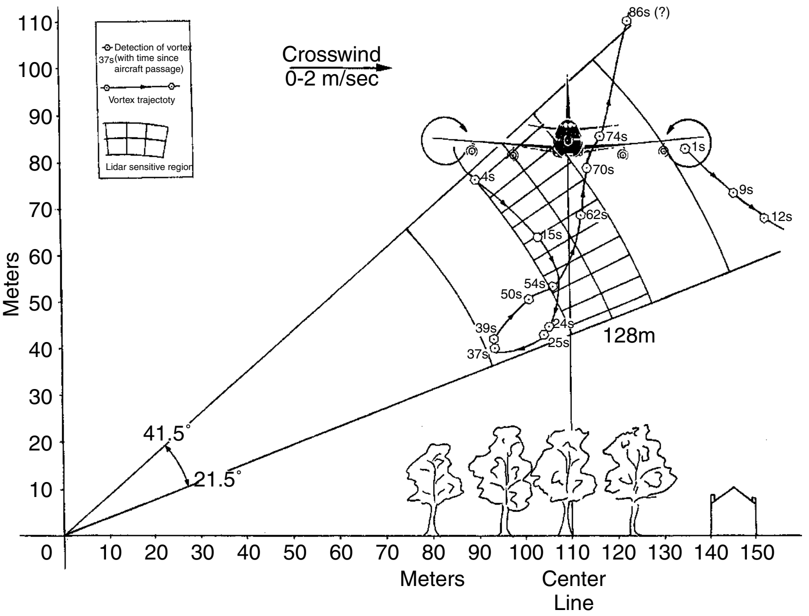

Detection of vortex
37s (with time since aircraft passage)
Vortex trajectoty
Lidar sensitive region
Crosswind
0-2 m/sec
86s (?)
74s
1s
70s
9s
12s
4s
62s
15s
54s
50s
39s
24s
37s
25s
128m
41.5°
21.5°
Meters
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
Meters
Center Line

Figure C3.2.26. Reconstruction of vortex trajectories for record CAA 80039 (figure C3.2.26). The lidar was sited at a position ~110 m north of the centre line with the aircraft (B747-200) ~85 m above ground flying E to W. The lidar was set to focus at ~128 m range and the scan in the N–S plane is shown extending over 31.5 ± 10°. The wind was WNW, about 5–6 ms^{-1} at ~295° at a height of 75 m, giving a variable cross wind of 0–2 ms^{-1} from the north. The weather was cold (3–5°) with occasional rain and sun.

transport aircraft now being developed (e.g. the Airbus A380). For airport operation, two types of capacity improvement may be envisaged:

- *Strategic*: scheduled increase in arrival/departure slots. Increased knowledge of vortex decay and environmental interaction could lead to small capacity gains from refined separation standards. Even a few slots per day at a busy airport would be of great value.
- *Tactical*: considerable scope for local hour-by-hour ATC decisions to reduce separations based on meteorological and lidar (LDV) information monitoring. Potentially large fuel savings and reduction of delays, particularly for aircraft in holding patterns, should be attainable.

This last point raises the question as to whether effective forms of routine monitoring by lidar could be developed to provide useful vortex advisory systems for air traffic controllers. Any such airport advisory system must of course be cost-effective to install and operate and hence the technology must above all be robust, reliable and easily maintained.

C3.2.6 Chemical, environmental and pollution study

The three primary phenomena for detection and measurement of chemical species and their environment have been noted in section C3.2.2: differential absorption, Raman scattering and fluorescence. Lidar techniques using these phenomena have been widely developed for studies ranging from short range monitoring (~10–100 m) across, e.g. chemical plants to measurements high into the mesosphere even above 100 km altitude. Differential absorption lidars (DIAL) are available commercially and may be employed routinely for pollution monitoring across, e.g. industrial sites and cities. Many lidars have been incorporated in high flying, long-range aircraft and have provided vital information on atmospheric chemistry and dynamics, including analysis of polar ozone depletion. The basis of the techniques with a few recent examples are outlined. Detailed accounts may be found in, e.g. Refs. [18, 21, 22, 33, 97]. A concise review of DIAL and Raman lidar for water vapour profile measurements has been given by Grant [145] with 133 references.

C3.2.6.1 Differential absorption lidar (DIAL)

The simple principle of DIAL lidar is the comparison of the backscattered signal (from aerosols and molecules) for two laser beams traversing the same path; one beam is tuned to an absorption feature of the chemical species under study while the other beam is well off absorption. For a pulse lidar differential analysis of the two signals S_0 and S_{ff} with highly developed retrieval algorithms [146] will provide a range resolved concentration density map of the absorbing chemical species. For some specialist applications, cw lasers operating on fixed paths to topographic targets or retroflectors can provide a path-integrated measurement. This may be suitable for, e.g. 24-hour surveillance across a chemical processing plant with the beams successively traversing various sensitive areas of processing reactors, storage tanks, etc. Such lidars can guard against any sudden catastrophic escape of materials and contribute also to the detection of low level leaks with improved operating efficiency.

A vast range of lasers have been used for DIAL equipments ranging in wavelength from near UV to mid-IR. Typically in the range 250–435 nm, species such as toluene, benzene, SO_2, NO_2 and ozone can be detected with dye lasers or frequency doubled or tripled light from, e.g. a Ti sapphire laser. Water vapour and molecular oxygen have absorption features with strong, well resolved lines in the near-IR in the region 720, 820 and 930 nm (water vapour) and 760 nm (oxygen). Dye lasers, tunable vibronic and solid-state lasers have been employed amongst others. The longer wavelengths 1–4.5 μm are well suited to detection of, e.g. methane, ethane, hydrogen chloride, hydrogen sulphide, nitrous oxide and many hydrocarbons. Varying laser dyes with sum and frequency difference mixing in non-linear crystals provides tunable coverage of much of the range, together with tunable OPO lasers. In the range 9–12 μm, covered by CO_2 laser wavelengths, there are many absorption features for hydrocarbons and various pollutant and toxic materials which can also be investigated by DIAL.

A schematic of a DIAL lidar is shown in figure C3.2.27; clearly, a basic technical requirement is the accurate positioning of the laser emissions with respect to the selected molecular absorption line (typically within $0.001\,\mathrm{cm}^{-1}$), with the laser line width narrower than the absorption feature. Separation of the scattered radiation at the two wavelengths (if transmitted simultaneously from two laser sources) may be achieved with narrowband filters. Otherwise, switching between the two 'on' and 'off' lines must be accomplished swiftly and precisely with typical pairs of pulses at >10 Hz rate. Attainable accuracy for concentration measurement depends on the parameters of the specific application, e.g. spatial and temporal resolution and the material under study. Typically detection limits below 10 ppb can be achieved with an accuracy of better than a few ppb. Measurements in the atmosphere for meteorological investigations have been widely developed. With molecular oxygen as the absorber and operating in the wings of the absorption lines, surface pressure and pressure profile can be measured with an accuracy of order 0.2%. Similarly, the temperature profile may be measured with an accuracy of typically better than 1 K through the tropopause using highly excited vibration lines of molecular oxygen with appropriate choice of emitted laser frequencies. Many intercomparisons and sensitivity analyses of different equipments have been made (see, e.g. Refs. [147–149]).

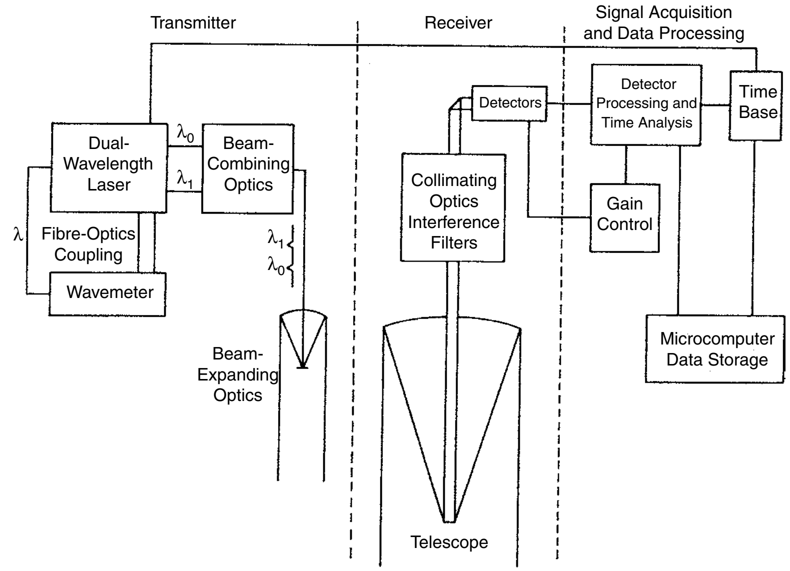

Transmitter
Receiver
Signal Acquisition and Data Processing
Dual-Wavelength Laser
λ_0
λ_1
Beam-Combining Optics
λ
Fibre-Optics Coupling
Wavemeter
λ_1
λ_0
Beam-Expanding Optics
Detectors
Collimating Optics Interference Filters
Telescope
Detector Processing and Time Analysis
Time Base
Gain Control
Microcomputer Data Storage

Figure C3.2.27. Schematic of a DIAL lidar showing the laser's transmitter and receiver.

For the shorter wavelengths, techniques of direct detection are employed with narrow-band optical filters for isolation of the required wavelengths and reduction of background. At longer wavelengths, heterodyne techniques have been investigated in order to improve SNR with inherently noisy detectors. The speckle-induced fluctuations in such coherent systems require averaging, but useful improvements in range performance with CO_2 DIAL systems have been demonstrated (see, e.g. Refs. [150–152]). Speckle correlation and averaging with two or more wavelengths for improved accuracy in DIAL have also been examined [153].

DIAL systems have now been used for many years in both fixed and mobile ground stations. An interesting comparison of results for two high-performance ozone DIAL lidars is shown in figure C3.2.28 (from Ref. [148]). A number of extensive campaigns have been conducted across European cities

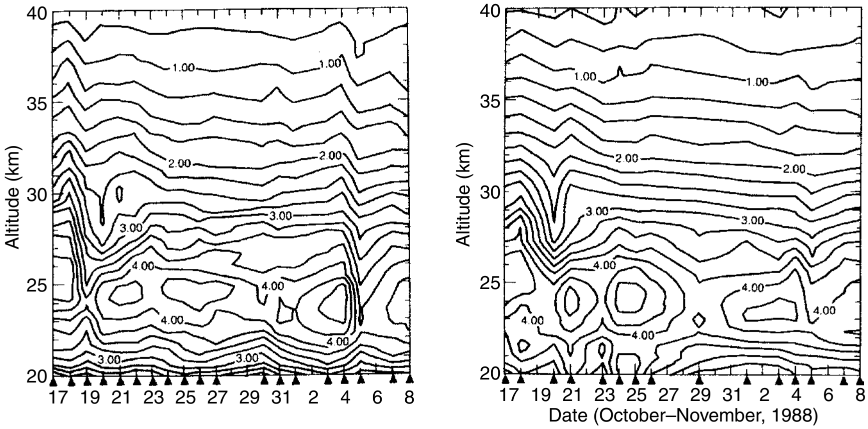


Figure C3.2.28. Comparison of the results for two ozone DIAL lidars spaced close together. The vertical arrows show the measurement day during the trial's period [148].

(e.g. Athens in 1994, Grenoble in 1999) bringing together a range of lidars and monitoring equipment for study of ozone smog and pollution episodes. Several large permanent ground stations (e.g. at Eureka in the Canadian High Arctic [35], at Alomar in northern Norway [112] and at OHP in southern France [154]) have been established for longer term monitoring of the atmosphere, often combining different lidar facilities with DIAL, Rayleigh, Mie and Raman lidar detection channels.

Measurements from aircraft have been particularly notable, particularly in the combination of aerosol and DIAL lidars. In the winter of 1999–2000, for example, the NASA Langley airborne equipment measured ozone at DIAL wavelengths 301.6 and 310.9 nm and aerosols at 1064, 622 and 311 nm, giving vertical profiles from 12 to 28 km across the wintertime Arctic vortex. Further valuable information on the interaction of PSCs and their implications for chemical depletion of O_3 was obtained [22, 33, 155].

C3.2.6.2 Raman lidar

Raman scattering provides an effective means for chemical species and aerosol measurement with lidar systems operating at shorter wavelengths to give the strongest scattering. Frequency doubled or tripled Nd-YAG lasers have been used; with an XeF examiner laser operating at 315 nm, Raman returns from O_2 (371.5 nm), N_2 (382.5 nm) and water vapour (402.8 nm) are created. Systems typically operate at several hundred Hz with 50–100 mJ per pulse, with collecting telescopes up to 1 m diameter and range resolution 20–100m. The role of water vapour in atmospheric models for numerical weather prediction (NWP), and as the leading greenhouse gas, is particularly important and Raman lidar measurements offer valuable information (see, e.g. Ref. [156]); a measurement example is shown in figure C3.2.29 [157]. Calibration techniques for Raman lidar water vapour measurements have been widely investigated

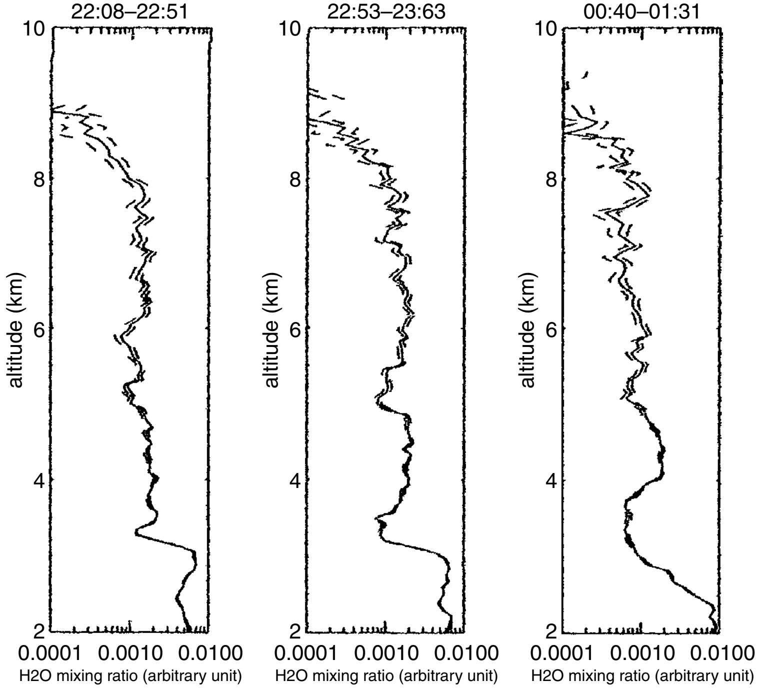


Figure C3.2.29. Example of Raman lidar measurement of water vapour in the atmosphere [157].

(e.g. Ref. [158]) using, for example, comparison with radiosondes providing local humidity data (see, e.g. Ref. [159]) and pressure and temperature data at cloud base with expected 100% humidity (see, e.g. Ref. [160]). Cloud liquid water content has also been an important study [161] for Raman lidar; liquid water has a broadened spectrum ~7 nm wide shifted about 5 nm from the narrow vapour return. Interference filters ~4 nm wide readily separate the two signals.

Another important application of Raman lidar has been in measuring aerosol extinction [162, 163]. By measuring the Raman N_2 signal at the shifted wavelength and the backscatter signal (molecular and aerosol) at the laser wavelength, the height-resolved aerosol extinction σ_{aer} and the extinction to backscatter ratio $\sigma_{\mathrm{aer}}/\beta(\pi)_{\mathrm{aer}}$ can be determined (see, e.g. Ref. [164]).

C3.2.6.3 Fluorescence lidar

Fluorescence lidar can be applied in several areas: for detection and identification of organic and biological materials on surfaces (by land or water), for detection of fluorescent dye particles (FDP) released as tracers in the atmosphere and temperature and concentration measurement of atoms at very high altitudes.

For organic and biological materials, two diagnostics may be employed. First, the fluorescent wavelength edge above which (as the laser source is tuned to longer wavelengths) the material no longer exhibits fluorescence. Second, there is the characteristic (usually quite broad band) spectrum of the fluorescence itself. The technique has been applied to pollution studies of, for example, oil spillages (accidental or deliberate) at sea with the possibility of large area, rapid surveillance by aircraft. Similarly the technique may be applied to biological studies of water—certain algae and simple life forms have characteristic fluorescence spectra. Early detection and identification can be achieved and monitoring of development as, for example, in the phenomenon of algae blooms. The use of FDP tracers in the atmosphere has been described by Uthe [165] (see also Ref. [166]). Organic resin particles of ~2 μm diameter containing a fluorescent orange dye have been used. In one example, a parcel released at 2.3 km altitude was tracked (by airborne lidar) for over 8 h and 200 km passage [165].

The first lidar measurement of sodium high in the mesosphere was made as long ago as 1969. This region is in fact difficult to study by other techniques, being inaccessible to balloons, aircraft and most satellites. It has many interesting features containing the coldest part of the atmosphere—the mesopause, metal layers formed by meteor ablation and polar mesospheric clouds (PMCs). Knowledge of the temperature, density and their variability (as a function of season, latitude and solar cycle, etc) is important for many geophysical phenomena. Several fluorescence lidar have been built variously operating at resonance wavelengths for Na, Ca, K and Fe. Typically, a range of dye lasers pumped by Nd-YAG have been used with pulse energies in the range 5–100 mJ at typically ~20 Hz. All solid-state generation of the Na D_2 resonance line at 589 nm has also been achieved by sum-frequency mixing Nd-YAG 1064 and 1319 nm radiation with lithium niobate or triborate non-linear crystals.

For relative concentration measurements, direct detection may be employed with narrow band filters (typically ~1 nm) to reduce background light. It is also possible to simultaneously measure density and temperature and a number of techniques have been developed [167–169]. For sodium, two or more laser frequencies are transmitted in turn, one exactly at the peak of the Na D_2 line and the other shifted 800 MHz by an acousto-optic modulator. The relative scattering at the two wavelengths gives a measurement of the temperature. An example of temperature and Na density profiles in the altitude range 80–105 km is shown in figure C3.2.30 [170]. For iron, the technique relies on the temperature dependence of the relative scattering for two closely spaced Fe atomic transitions at 372 and 374 nm. These two lines originate on the lowest atomic ground state and an upper sub-ground state about 416 cm^{-1} higher; their population and hence the ratio of scattered light is determined by Maxwell–Boltzmann statistics in thermal equilibrium. Generation of the required radiation has been achieved

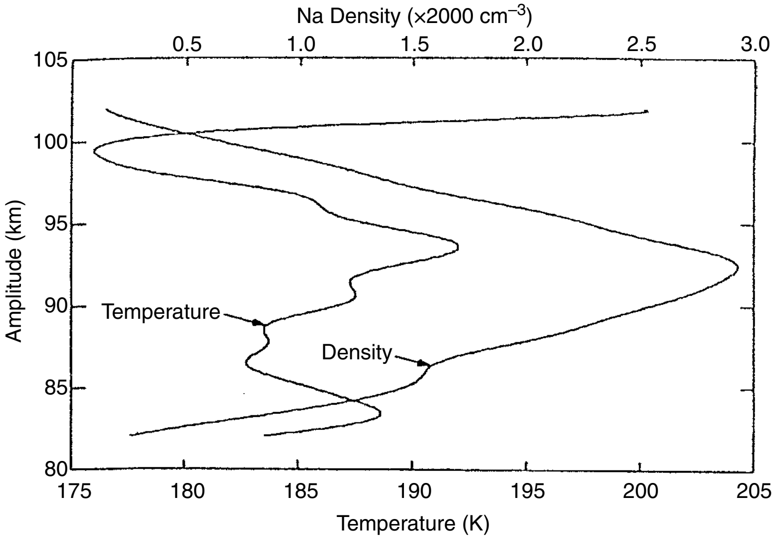


Figure C3.2.30. Example of fluorescence lidar measurements of temperature and Na density high in the atmosphere [170].

with independent, frequency-doubled and injection-seeded pulsed alexandrite lasers. Several equipments, ground and airborne, have now been built for mesospheric studies with routine measurements over extended periods, and more concentrated trials for expected meteor showers. Figure C3.2.31 shows a record of such an event [171].

C3.2.7 Lidar and military applications

> *"Many think that in some way they (the Martians) are able to generate an intense heat in a chamber of practically absolute non-conductivity. This intense heat they project in a parallel beam against any object they choose by means of a polished parabolic mirror of unknown composition…"*
> *H G Wells: The War of the Worlds (1898)*

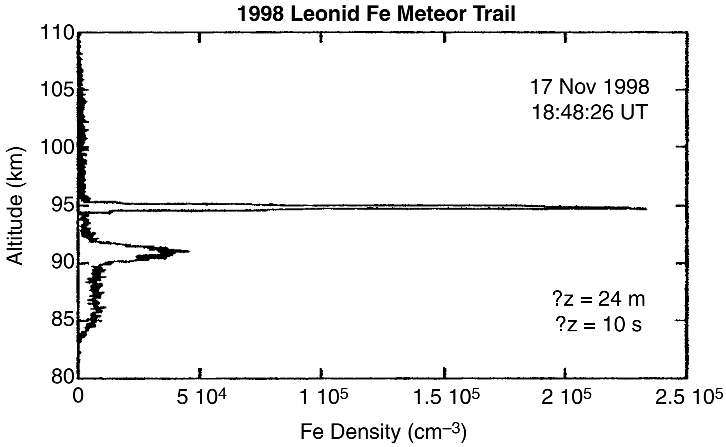


Figure C3.2.31. Record of a meteor ablation trail made with a Fe resonance lidar [171].

C3.2.7.1 Introduction

In the popular mind, lasers probably represent the ultimate science fiction offensive weapon. After much media hype through the late 1980s, fuelled by the United States' Strategic Defence Initiative (popularly known as 'Star Wars'), the injection of very large funding and remarkable computer graphics, it is true, for example, that a leading US defence company claimed in 1996 to be able to build an airborne laser cannon that would destroy enemy missiles at ranges of several hundred kilometres. This perhaps neatly concludes the requirement, allegedly laid down in the 1930s by the British War Office to early proponents of radio waves, that their 'death ray' should be capable of killing sheep at a few miles.

Just as such early studies led to radar and its enormously important impact on the Second World War, the remote sensing capabilities of lasers are now becoming increasingly important. In fact, it may be said that later sensing is one of the few truly new (non-nuclear) military techniques introduced since the Second World War. Even modern thermal imaging and its immense capabilities had its genesis in the basic work and crude systems introduced in the late 1930s (see, e.g. R.V. Jones, *Most Secret War*, 1978).

In reviewing laser radar applied to military tasks, one can distinguish three successive phases. The first phase consists of relatively simple, direct-detection methods for very successful range finding, target marking and laser beacons employing near-IR ruby and Nd-YAG lasers. The non-eye-safe character of these laser systems gives problems of troop training in realistic conditions, and this has led in a second phase to the introduction of longer wavelength lasers such as CO_2 and Ho-YLF for basically similar tasks. In these phases, the laser is not much more than a bright light source. The third phase, which employs both advanced direct-detection and coherent techniques and exploits the full laser capabilities for wind measurement, target identification, chemical sensing, imaging, etc, is potentially of considerable military significance in the 21st century. Before outlining this work, however, it is worth reviewing the strengths and weaknesses of lidar in comparison with other military sensing technologies.

In many ways, lidar is very complementary to thermal imaging. In particular, CO_2 lasers at $10\,\mu m$ wavelengths lie in the middle of the $8-13\,\mu m$ band of longer wavelength thermal imagers, and thus their atmospheric penetration and range performance is likely to be very similar. On the other hand, passive thermal imagers, with rapidly scanned multi-element detectors, provide excellent surveillance over quite large fields of view. In contrast, lidar, with few or single-element detectors and the very precise pointing capabilities of a diffraction limited laser beam, seems better suited to detailed interrogation and investigation of small regions identified as 'interesting' by other means. Compared with longer wavelength centimetric radar, lidar has considerably poorer transmission, at least in the lower atmosphere, and thus most battlefield-type tasks have been confined to tactical ranges of a few tens of kilometres at most. This, of course, no longer holds true high in the atmosphere, and it is also worth appreciating the limited penetration of millimetric radar through, for example, heavy rain (due to large scattering from millimetric-size rain drops). Covertness is another important aspect; it is quite difficult for a potential target to detect that it is being illuminated with a comparatively low power, often short-pulse, laser beam, and much attention has been devoted to this. However, radar, unless used with great discretion, has been described as a beacon signalling to its adversary 'here I am—come and hit me'. Finally, it is worth pointing to a few unique military capabilities of lidar; these include, for example, wind sensing, detection of very small 'targets' such as wires and cables and remote chemical detection.

C3.2.7.2 Range finding, target marking, velocimetry, etc

Simple rangefinders based on short-pulse, flash-lamp pumped, solid-state lasers (ruby, Nd-YAG) were amongst the first developments of military hardware. Using the elementary timing principle (section C3.2.2.1, figure C3.2.2*c*), very compact, robust and efficient (often battery-operated) systems were rapidly instituted at the infantry rifleman level, for tank rangefinders and also for aircraft operation. In a clear atmosphere, reliable operation over several kilometres range is easily obtained with a few metres accuracy corresponding to ~10 ns pulse length. In fog and smoke, longer-wavelength lasers generally have superior performance with the added advantage of eye safety. Over the years, very extensive studies have been carried out on relative effectiveness under varied conditions of humidity, fog, smoke, camouflage, target characteristics, etc, and laser ranging systems of many different levels of sophistication have been developed. For example, scattering from a smoke-screen laid down in front of a target is likely to give a strong signal similar to that from the obscured target itself (supposing that the laser beam can adequately penetrate the smoke). Logic that selects this later pulse signal from the target rather than the smoke signal is required (figure C3.2.32*a*).

In the related technique of target marking, the selected target is illuminated with a laser. Scattered radiation may then be picked up by a lidar receiver housed in a guided bomb or missile and provides information that directs in onto the target (figure C3.2.32*b*). These methods showed their effectiveness in the 1970s in the Vietnam War; many unsuccessful attempts had been made to destroy an important

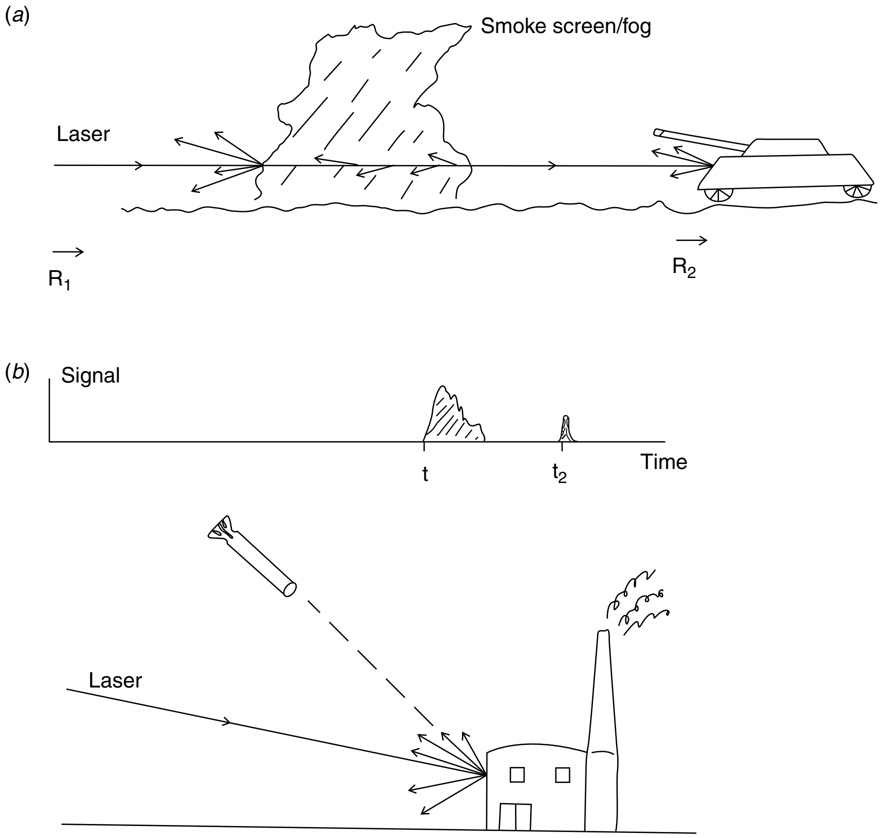


Figure C3.2.32. (*a*) Schematic of a laser rangefinder with signals from different ranges. Last-pulse logic is required to select the correct return from the target. (*b*) Schematic of guided missile homing on light scattered from the target marked by an incident laser beam.

bridge by conventional bombing which was eventually knocked out first time by a laser-guided device. More recently, the sophisticated development of such laser-guided missiles (LGMs) first impacted on the popular consciousness with extensive newsreel coverage of missiles apparently penetrating ventilation shafts, bunkers, etc, during the 1991 Gulf War.

For rangefinding, the use of a short pulse is but one extreme of coding the transmitted radiation. Many other forms of code may be used to acquire range or equivalent time-delay information, just as in longer-wavelength radar. These include amplitude modulation (AM) with, for example, a pseudo-random on-off code; auto-correlation of the strength of the return signal with the AM code itself gives it a peak response at a time interval t_f corresponding to the range to the target and back. Alternatively, frequency modulation (FM) of the transmitted beam may be employed and has the potential advantage of providing Doppler velocimetry information about the target, as illustrated in figure C3.2.33*a* with a linear chirp pulse of duration Δt_C and frequency modulation f_C. The return signals of a stationary target and moving target (with negative Doppler shift f_D) are shown. Signal-processing techniques of pulse compression ensure that the range resolution in this case is not determined by pulse length Δt_C, but rather by the rate of change of frequency (the chirp rate). Some compromise of range and Doppler frequency resolution is required but, typically, 10 m range resolution and 1 m s^{-1} velocity resolution has been demonstrated in reasonably compact coherent systems at ranges up to 10 km. The sign of the Doppler shift may be established by use of the successive up and down chirps, as illustrated in figure C3.2.33*a*.

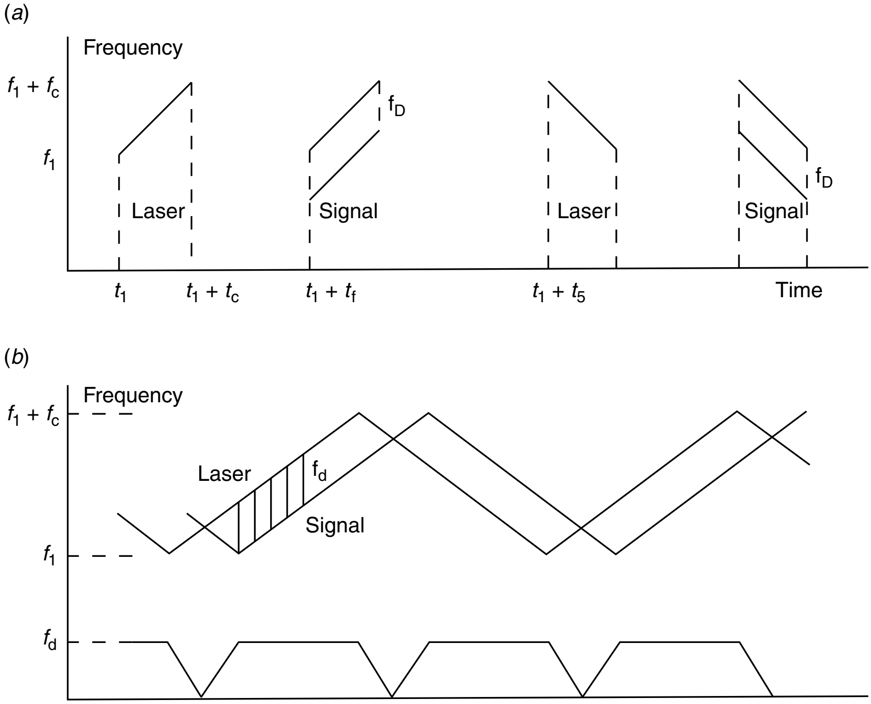


Figure C3.2.33. Coding of laser beams for simultaneous range/Doppler measurement. (*a*) Up/down linear chirp pulses of duration Δt_c, separation t_s and frequency modulation f_c. (*b*) Frequency modulation of a continuous wave beam (FMCW). For a distant stationary target, the range is proportional to the frequency difference f_d of the outgoing beam and return signal as shown.

Frequency modulation of a continuous wave beam (FMCW) may also be used, as illustrated in figure C3.2.33*b*. With the delayed return signal beating against the outgoing waveform as shown, simple spectral analysis gives a frequency shift proportional to the range to the distant (stationary) target. For a moving target, this splits into two frequencies separated by $\pm f_D$. In practice, careful consideration of the various parameters indicated in figure C3.2.33*b* of FM, code length and the target properties of range, velocity, etc, are required, including the problem of range-Doppler ambiguity. Nevertheless, several systems have been developed demonstrating high performance for simultaneous measurement of range and target velocity.

C3.2.7.3 Micro-doppler and vibrometry

The bulk motion of a distant object may provide useful military information, as is demonstrated in many combined rangefinders and Doppler velocimeters. In addition, the information due to very small movements and vibrations within an object may potentially be of value in both civil and military fields. For a simple surface vibrating sinusoidally with amplitude a and frequency f_v, the displacement $a(\sin 2\pi f_v t)$ gives a peak speed of $\pm 2\pi f_v a$. From equation C3.2.5, the peak micro-Doppler shift rises to $\pm 4\pi f_v a/\lambda$; with typical values of $a = 20\,\mu\text{m}$, $f_v = 200$ Hz, the peak speed is $\pm 5.03 \times 10^{-2}\,\text{m s}^{-1}$ and for $\lambda = 10.6\,\mu\text{m}$ the micro-Doppler shift is only ± 9.51 kHz. Such a shift is readily detected (to provide the amplitude) but, more importantly, techniques of frequency demodulation may be employed to give the underlying vibrational frequency f_v. The thrust of much recent work has been, on the one hand, to develop the lidar systems for such measurement (coping with objects which may also be in bulk motion) but, on the other hand, to establish how characteristic and reproducible the vibrational frequencies are from particular targets and whether they can provide a key to successful identification. It is well known, for example, that their precise blade rotational frequencies are specific to particular types of helicopter and provide a means of identification. In practice, the problem is extremely complex; within a vehicle, for example, the driving frequencies originating at the primary power sources of engines, electrical motors, etc, are coupled through all manner of nonlinear mechanisms to the outside panels and frames, etc, from which the laser radiation is scattered. These outside surfaces are likely to have their own resonant frequencies and are further influenced for a ground vehicle by the terrain—gravel, mud, tarmac, etc, over which it is proceeding. Nevertheless, some success is being achieved and has been discussed in open literature. At the very least, the state of readiness of distant vehicles and, for example, the difference between live, active tanks and dead, knocked-out tanks or decoys is detectable. As a further example, it should be possible to match the radar techniques of aircraft identification (from signals reflected from turbine blades within the jet engine itself) with lidar methods taking signals reflected from the outer airframe skin.

C3.2.7.4 Target imaging

The military lidar capabilities discussed to date of range finding, velocimetry and vibrometry may be considered as useful sensing discriminants in the broader task of providing a target image. These and various other discriminants available by electro-optic laser techniques are indicated in table C3.2.12.

The point is that, while one is accustomed to interpreting visual grey-scale or colour images, any discriminant that varies across a scene may be used to build an ‘image’ of that scene. This may itself provide valuable information about an object (the ‘target’) and a means of picking it out against the background.

Table C3.2.12. Remote sensing discriminants for creating an image of the target region.

Scattering reflectivity/intensity from strength of signal
Change of intensity at different wavelengths
Change of intensity at different polarisations
Occurrence of 'glint' signals – high intensity retro-reflections
Statistical character of signal fluctuations
Range as set in different range intervals
Doppler frequency shift – bulk velocity
Micro-Doppler – frequency modulation (FM) vibrometry
Signal Amplitude Modulation (AM).

Of the nine discriminants listed, the first five all relate to various aspects of the target region reflectivity and the signal intensity as potentially measurable in either a direct detection or coherent lidar. Under laser illumination, the scattering characteristics of different surfaces—vegetation, metal, paint, concrete, soil, grass, camouflage, etc—vary quite widely under changing conditions of illumination (angle of incidence, polarisation, wavelength, etc), as indicated in figure C3.2.34. The term 'glint' refers to the very strong, almost retro-reflection, signal that may occur on successive reflections from smooth surfaces (e.g. around a window frame or wheel arch) and is very characteristic of the presence of a man-made object as opposed to natural vegetation and soil. This is an extreme example of the difference of signal fluctuations as a laser beam is moved over different surfaces, e.g. concrete or brick compared with grass.

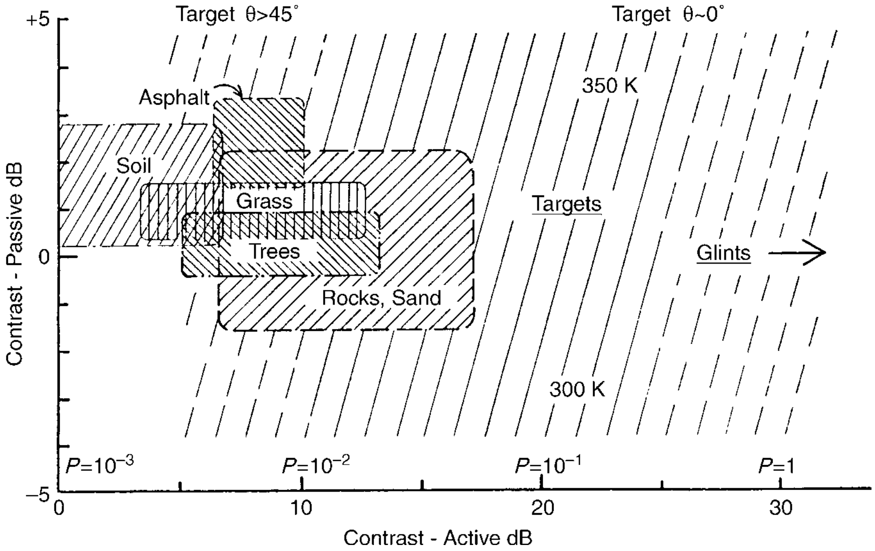


Figure C3.2.34. Example of the active laser radar (10 μm) and passive thermal imager (8–12 μm) contrasts to be expected between various targets and standard background (supposed at 300 K) temperature emissivity 0.999 and reflectivity $\xi = 10^{-3}$). The large variation can be attributed to different angles of incidence, degree of moisture, etc (from Ref. [191]).

These and the other four discriminants listed in the table—range, bulk velocity, vibration and any amplitude modulation (indicative of moving machinery for example)—may be used to create and active ‘image’ of the target region with suitable colour coding, highlighting of significant regions, etc. The pixel size or spatial resolution in the image will be determined by the usual optical parameters—the diffraction limits and atmospheric refractive turbulence—and at 10 μm should provide a good match to co-located thermal imagers surveilling the passive scene. However, it is worth noting the prime military imperative of speed of measurement and the impact that has on optical scanning arrangements and the lidar dwell time on each pixel element. For passive imagers, whether thermal or visual, all pixel elements in the scene are radiating/reflecting energy simultaneously and the angle of view and overall speed of measurement is ultimately determined by the number of individual elements (each viewing one pixel) that can be packed into the detector. With an active lidar system, maximum sensitivity is achieved by illuminating each pixel sequentially with the laser beam, making the measurement and moving on to the next pixel. Necessarily, this is likely to be a slower process and reinforces the utility of lidar for provision of unique information from detailed interrogation of small regions as opposed to large area surveillance. The alternative of floodlighting a region, across many pixels, with a high-power expanded laser may be feasible in some circumstances but obviously mitigates against low-cost, low-power requirement and covertness.

A number of lidar images are shown in figure C3.2.35 and illustrate the potential. As always, extensive deployment into the armed forces requires that such systems are made cheap, compact, reliable and easy to operate and maintain.

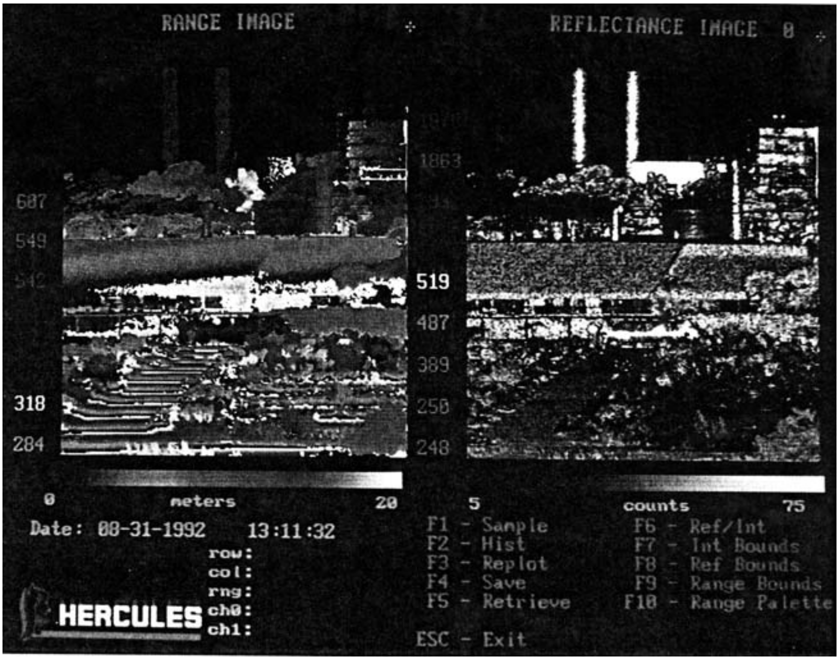

RANGE IMAGE
REFLECTANCE IMAGE 0
687
549
542
318
284
1870
1863
519
487
389
250
248
0
meters
20
5
counts
75
Date: 08-31-1992 13:11:32
row:
col:
rng:
ch0:
ch1:
HERCULES
F1 - Sample
F2 - Hist
F3 - Replot
F4 - Save
F5 - Retrieve
F6 - Ref/Int
F7 - Int Bounds
F8 - Ref Bounds
F9 - Range Bounds
F10 - Range Palette
ESC - Exit

Figure C3.2.35. Range and reflectance imagery with a 1.047 μm direct detection lidar at 2 km, collected with the Hercules Corporation system (from Ref. [192]). Much detail is lost in transcription from the original false colour imagery.

C3.2.7.5 Wind measurements: rocketry, landing aids, ballistic air drops

As described in section C3.2.4, rapid measurement of wind field with good spatial resolution at ranges between a few tens of metres to many kilometres is unique to laser radar. In consequence, it is not surprising that such capabilities have been extensively explored for military tasks. An immediately obvious area is the impact of wind on accuracy of munitions. Rockets, in particular during the early stages of flight, are travelling comparatively slowly and may be quite strongly influenced by the local wind conditions. For example, measurement of flow and downwash under helicopters has been measured by lidar in order to improve delivery point accuracy of rockets fired from such platforms. Very large ground-based rocket systems are also influenced by the wind; some attempt to allow for this may be made by applying a correction in the fire control system. Conventionally, this may be a forecast wind or from a distant radiosonde measurement, possibly made several hours and tens of kilometres away. With the vagaries introduced by local terrain and natural wind fluctuations, the remaining uncertainties may contribute appreciably to the budget error for the weapon. For example, a compact coherent CO_2 lidar was developed in the late 1980s for the US Army Multiple Launch Rocket System (MLRS). The cw lidar would map the wind field at a distance of 100 m and it was said it could be produced, in volume production, at a competitive cost.

The potential of wind measurement as a landing and take-off aid has also been described. This could assist the landing of helicopters and fixed-wing aircraft on ships with monitoring of the complex wind flows around the landing deck. Such systems could also provide real, full-scale measurements around such structures and, together with wind tunnel testing and CFD modelling, assist in improved ship and platform design.

An airborne lidar system that profiles the wind field below the aircraft has recently been described. Parachute delivery of cargo and conventional bomb drops without guidance control are obviously much affected by the prevailing winds. If these are known, the release point can be adjusted to compensate. However, such information may not be available or easily attainable in remote or hostile locations. A 2 μm lidar system based on a flashlamp-pumped Cr, Tm:YAG laser was built by Coherent Technologies and flown from the USAF Wright Laboratory. The Q-switched laser gave 50 mJ pulses at 7 Hz repetition rate and the system was easily installed in a C1411 aircraft. The scanner mirror mounted on a wedge produced a 20° × 28° (half angle) elliptically conical scan pattern with minor axis along the aircraft's longitudinal axis. Signal processing limited the first test in June 1995 to a range of 4.3 km, equivalent to a vertical distance of 3.8 km (15 000 ft) with the scan angles employed. Excellent agreement in the comparison of wind profiles measured with the airborne system, two ground-based lidars (one 2 μm and one CO_2 10 μm) and a radiosonde was found. Further studies assisted in a high-altitude bombing test above 30 000 ft (8 km) by B-52 aircraft. Using a composite wind profile, the bombers could correct for the wind and substantially improve their drop accuracy.

C3.2.7.6 Chemical detection

The DIAL technique described in section C3.2.6 suggests the possibility of detecting the poison gas and toxic nerve agent materials that might be used in chemical warfare (CW). The problem is that such materials may be chemically quite complex with rather broadband regions of absorption and transmission that makes positive discrimination and identification difficult. This contrasts with the detection of simple atmospheric constituents and pollutants with generally narrow absorption lines, as discussed in section C3.2.6. In the latter case also, the lidar would usually be built with the aim of detecting certain well-defined chemical species; in practical operation, the equipment would be set to concentrate on very specific materials. In the military case, however, a range of toxic materials might be present and the lidar must be capable of seeking and detecting all of them without any prior indication of which is actually there. Nevertheless, the DIAL technique has been seriously examined and the US army at least has developed a CO_2-based chemical detection system. With isotopic lasers, the overall tuning range may extend from 9.1 to ~11.8 μm, which offers a reasonable band of frequencies for discrimination in the longer-wavelength region. Solid-state lasers based on non-linear optical materials may also offer broad tuning ranges in wavelengths up to ~5 μm.

Finally, it is worth noting that some attention has been given to the possible remote detection by lidar of materials used for biological warfare (BW). In this case, discrimination against similar materials (pollens, spores, etc) occurring naturally in the atmosphere would seem very difficult. However, it has been suggested that an airborne lidar searching for anomalies and discontinuities in the atmospheric aerosol scattering (section C3.2.4) might provide an indication of the possible release of such BW material at several tens of kilometres range.

C3.2.7.7 Terrain following and obstacle avoidance

Aircraft flying at moderately low level are extremely vulnerable to fast, agile missiles fired from the ground. One solution is to fly at even lower level to provide an adversary with less warning of approach. Such 'nap-of-the-Earth' operation increases the obvious dangers of failing to clear the ground or flying into natural or man-made obstacles. Laser radar with its capability for very rapid measurement, fine pointing and range resolution has the potential to provide a valuable navigation and obstacle warning aid, as illustrated in figure C3.2.36. Several lidar systems have been developed worldwide to explore this potential.

Figure C3.2.37 shows a terrain map generated by a CO_2 lidar developed in France and flown on a Puma helicopter. In this case, the lidar was an FM/CW coherent system with the beam directed into a pair of rotating prisms that generated a scanning rosette pattern. This pattern was composed of 4000 dots or pixels, each one corresponding to a pulse from the signal processing. The maximum scanning

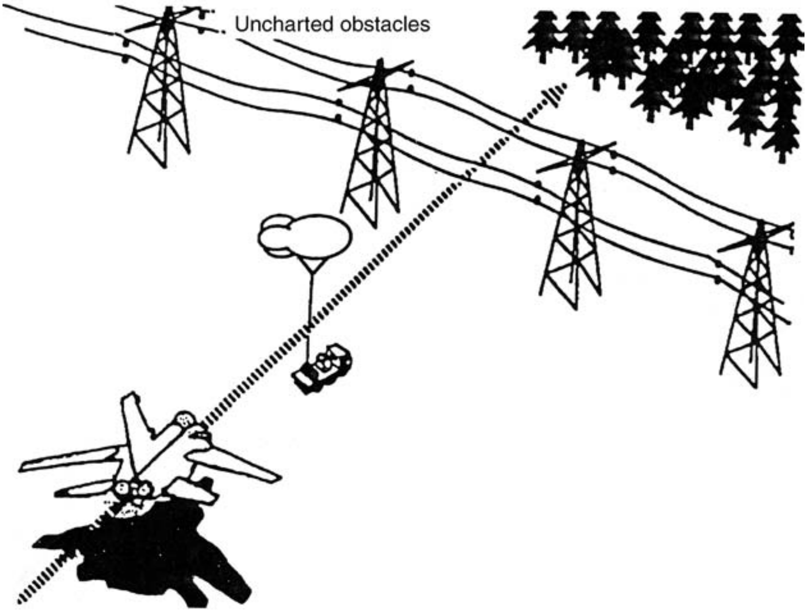


Figure C3.2.36. Potential obstacles in the path of a very low-flying aircraft that might be detected by a rapidly scanning laser radar.

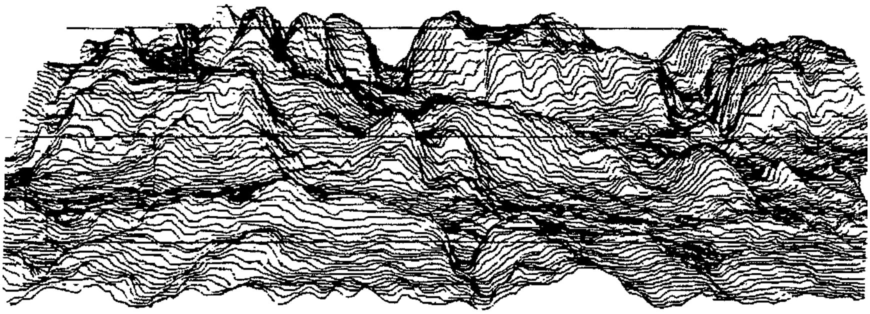


Figure C3.2.37. A terrain map, 10 × 7 km generated by a scanning laser radar (coherent FMCW CO_2) flown on a helicopter (from Ref. [193]).

rate was ~2.5 Hz. An Anglo-French technical demonstrator programme was developed with acronym CLARA (Compact Laser Airborne Radar). This advanced 10 μm pulse heterodyne system was trialled on both a fixed-wing Tornado aircraft and a Puma helicopter and followed the earlier LOCUS (Laser Obstacle and Cable Unmasking System) developed by two groups within the GEC Marconi Company and flown on A6-E, HS748 and Tornado aircraft. As indicated in figure C3.2.38, the technology to detect power cables and display them to the aircrew was demonstrated. In addition to the two primary roles of obstacle warning and terrain following, CLARA was also designed for targeting and true air speed measurement. A pod-mounted configuration was adopted with complex optical scanner assembly to address several quite different modes of operation. Detection, classification and real-time display of

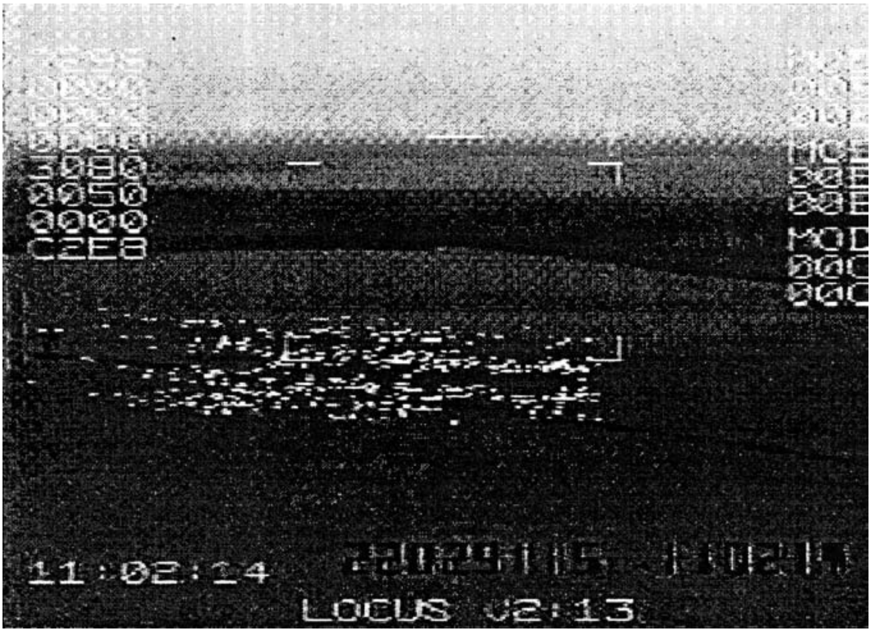


Figure C3.2.38. Taken from a video record made with the LOCUS CO_2 pulse heterodyne lidar, showing the lidar returns from cables. Much of the detail is lost in reproduction (from Ref. [194]).

obstacles, including various cables, masts, buildings and trees, must be achieved in daylight, at night and in adverse weather. In order to ensure that suitable warnings continue to be provided when the aircraft is turning, a large field of regard, the size of which is governed by the flight envelope of the host aircraft, must be adopted for the sensor.

C3.2.7.8 Star wars

In the original Strategic Defence Initiative (SDI), it was envisaged that lasers would have a central role, both as weapons for destroying enemy missiles and as sensors of such missiles. An important aspect of such sensing is to distinguish between decoy missiles and those actually containing warheads. A number of investigations in the United States have been reported in the open literature of the technical problems of laser pointing, tracking and examining high-flying rockets and space vehicles (see, e.g. Ref. [172]). Figure C3.2.39 gives an artist's conception of the large installation of the Firepond Laser Radar Research Facility of the Lincoln Laboratory where much of this work has been done. In this equipment, several laser radars, but principally a large CO_2-pulsed coherent system, were used in a succession of long-range measurements. In the early 1990s, the first range-Doppler images of orbiting satellites were collected at ranges of 800–1000 km and were later extended to 1500 km. The argon laser indicated in figure C3.2.39 provided the source for a visible light tracker of reflections from the satellites. Two initial SDI experiments were conducted later in 1990. These FIREFLY experiments were designed to investigate whether a laser radar could discriminate between a ballistic missile warhead and an inflatable decoy. In the tests shown schematically in figure C3.2.40, a rocket was fired from the NASA Wallops Island Flight Facility several hundred kilometres to the south, rising to an altitude of 560 km in a 750 s flight. Coherent CO_2 laser radar imaging was successfully conducted over the central 50 s of the flight at

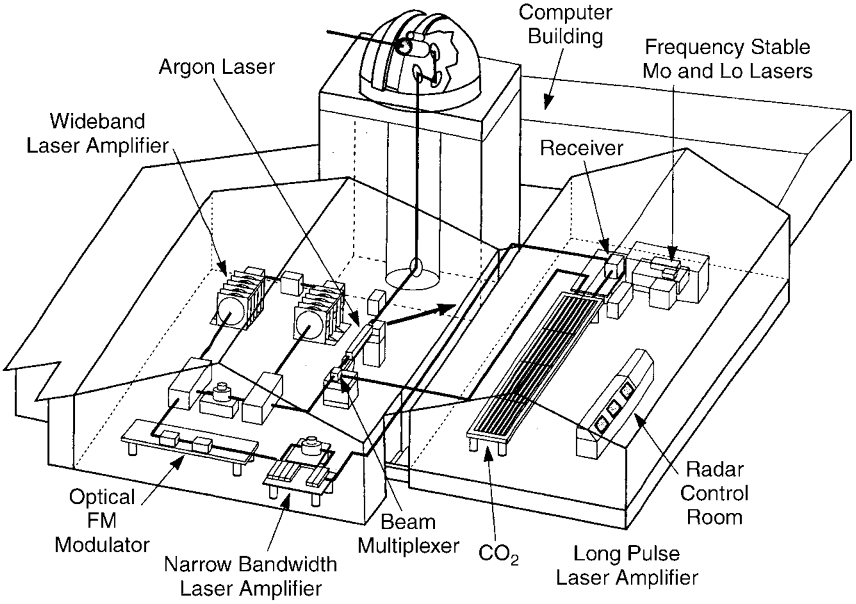


Figure C3.2.39. Artist's conception of the Firepond wideband imaging laser radar at the Lincoln Laboratory. Laser beams are transmitted from a 1.23 m diameter mirror Coudé telescope with central abscuration of 0.21 m (from Ref. [172]).

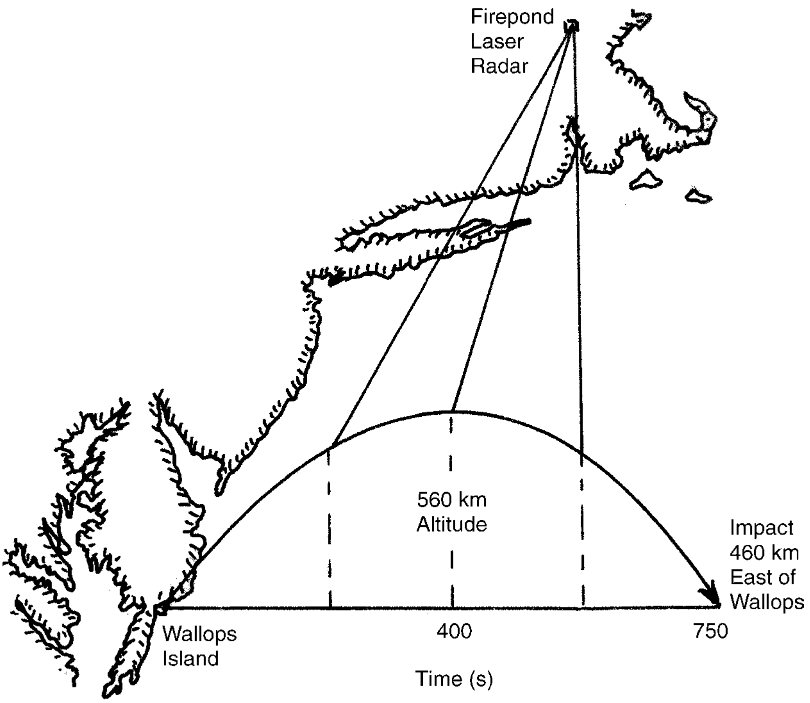


Figure C3.2.40. Schematic of the laser radar imaging during the SDI Firefly tests on missile detection in 1990 (redrawn from Ref. [172]).

a maximum range of 740 km. The prime target was a conical inflatable balloon made of carbon cloth. After approximately 360 s from launch, this was ejected in a small canister, was subsequently inflated and then set spinning in different manoeuvres to present a variety of target views to the imaging laser radar.

A second series of tests with acronym FIREBIRD (Firepond Bus Imaging Radar Demonstration) had the more ambitious objectives of investigating laser radar discrimination and countermeasure techniques. A high-performance booster rocket was used to deploy a dozen targets for study by several ground and airborne sensors along the eastern seaboard of the United States. These included passive IR sensors, UHF radars and optical sensors as well as the Firepond laser radars which also incorporated a photon-counting Nd-YAG laser radar. Generally excellent results were reported and completed the investigation of laser radar discrimination techniques. The advanced electro-optic techniques developed for this programme, including laser stabilization to the sub-Hertz level at high-power, precision coding of laser pulses, and sophisticated signal processing and analysis represent a remarkable *tour de force*. Demonstration of such precision tracking and imaging at over 1000 km ranges constituted at least a significant milestone in the history of laser radar.

C3.2.8 Lidar and geoscience

For the present purposes, lidar as applied to geoscience can be discussed in several broad areas: relatively short-range measurements to the sea surface and sea bed (few tens to hundreds of metres); sea, earth and ice surface from aircraft at up to ~15 km and finally ranging at hundreds to thousands of kilometres to satellite (and lunar) retroreflectors for studies of crustal dynamics, etc. These areas are outlined in turn.

C3.2.8.1 Airborne mapping of terrain and ocean shelf

Ranging to the Earth's surface has been touched on in section C3.2.7.7. Extensive studies with airborne systems have been conducted by several groups. Typical pulse energies of a few mJ with Nd-YAG lasers at 1.06 μm operating at up to a few hundred Hz have been employed. Figure C3.2.41 shows a block diagram of the lidar instrument developed at the Goddard Space Flight Centre [173] and mounted in a high-altitude aircraft; the position of the aircraft was measured to sub-metre accuracy by use of differential GPS receivers. Typical footprint size, from the beam divergence of 2.5 mrad, amounted to ~25 m at 10 km altitude. Problems of pulse waveform spreading and target signatures for such altimeters have been considered by several authors (see, e.g. Refs. [174, 175]). Reported investigations include study of craterform structures and observations of the Mount St. Helena Volcano [176, 177].

Laser depth sounding and the performance of a Swedish system has been well reviewed by Steinvall *et al* [178]. Several such lidars have been built worldwide, including, for example, in the USA [179], Australia [180] and also Canada, Russia and China. In typical operation from a helicopter, two pulsed laser beams are emitted simultaneously and scanned over the surface. The longer wavelength (usually 1.06 μm, Nd-YAG) is mostly reflected from the water surface whereas the shorter wavelength (green 0.53 μm, doubled Nd-YAG) better penetrates the water and may provide a return from the bottom layer. Such lidar bathymetry provides a most promising technique for high-density, rapid sounding of relatively shallow waters typically 1–50 m deep. In addition, its use can be established for other sensing applications, as noted in section C3.2.6.3—e.g. algae/chlorophyll monitoring and oil slick detection and classification. Operating from a helicopter also gives access to areas difficult for ships, e.g. around small islands, reefs and inlets. In addition, the rapid survey rate, typically 10–100 km^2 per hour depending on shot density, is particularly valuable.

In evaluation of performance, the prime question is the depth penetration, as largely determined by

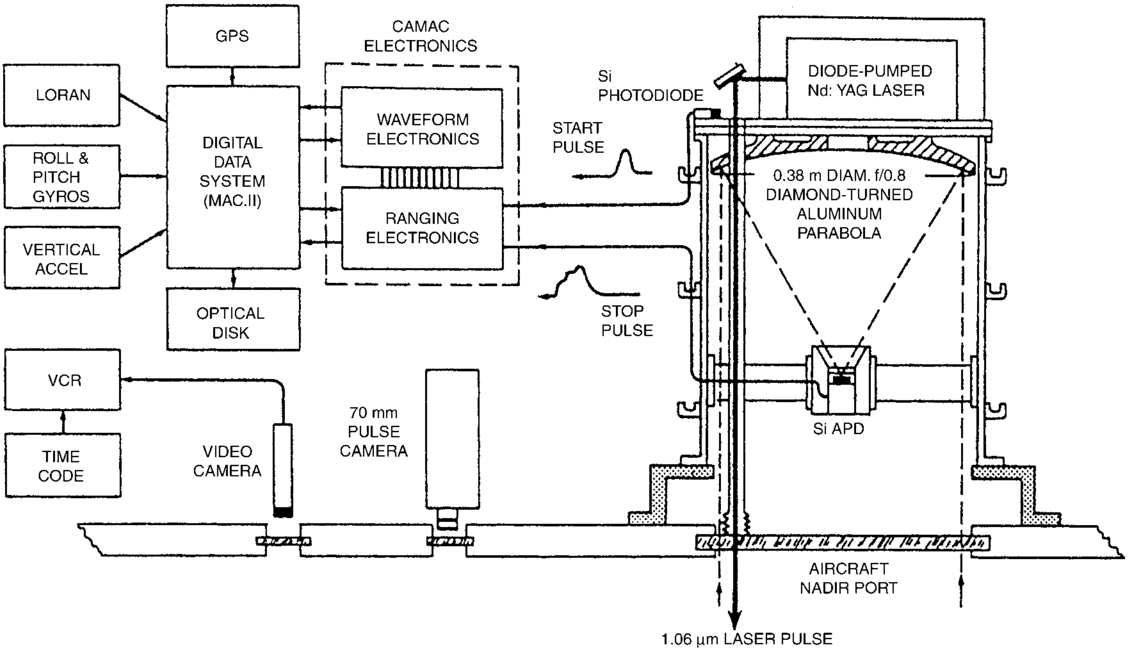

GPS
CAMAC ELECTRONICS
LORAN
ROLL & PITCH GYROS
VERTICAL ACCEL
DIGITAL DATA SYSTEM (MAC.II)
WAVEFORM ELECTRONICS
RANGING ELECTRONICS
OPTICAL DISK
Si PHOTODIODE
DIODE-PUMPED Nd: YAG LASER
START PULSE
STOP PULSE
0.38 m DIAM. f/0.8 DIAMOND-TURNED ALUMINUM PARABOLA
Si APD
VCR
TIME CODE
VIDEO CAMERA
70 mm PULSE CAMERA
AIRCRAFT NADIR PORT
1.06 μm LASER PULSE

Figure C3.2.41. Block diagram of the airborne laser radar built at Goddard Space Flight Centre [173].

water turbidity [181]. The maximum depth penetration D_{MAX} is given (see, e.g. Ref. [178]) by

$$D_{MAX} = \ln(P^*/P_{min})/2G \tag{C3.2.29}$$

with $P^* = S.\varsigma/H^2$, where S is a system parameter (derived from laser power, receiver aperture and loss factors), ς is the reflectivity of the seabed and H the lidar platform altitude. P_{min} is the minimum power for detection of the bottom of the sea (usually limited in daytime by solar background and at night by detector/electronic noise). The system attenuation G is a complex function of intrinsic water parameters—beam attenuation, scattering coefficients, etc. With typical lidar parameters, values of $GD_{MAX} \approx 3-6$ can be achieved. It should be noted that a large increase in system parameter S only gives a modest increase in depth performance D_{MAX} due to the exponential attenuation expressed in equation C3.2.29. Steinvall *et al* [178] gave the example of a factor 10 increase in S only adding to D_{MAX} by ~4 m for $G \approx 0.3\,\mathrm{m}^{-1}$ [$\log 10/(2 \times 0.3) \approx 4$].

Optical attenuation in sea water may be related to the Secchi disk depth D_s, which is the maximum depth at which a submerged white disk can be observed. Steinvall *et al* [178] showed that values of $D_s \approx 10$ m gave $D_{MAX} \sim 20-25$ m, and $D_s \approx 30$ m gave $D_{MAX} \sim 25-35$ m for a state-of-the-art lidar bathymeter.

Other important factors limiting system performance include strong winds (affecting helicopter flight path) and associated sea state (white and broken water reduces accuracy and beam penetration). Low-lying cloud and fog obviously inhibit performance and the thin haze (often quite dense and sometimes called ‘sea smoke’) above the water may provide a false return for the surface itself. Within the water, narrow scattering layers and other inhomogeneities may also provide spurious returns.

Finally, the nature of the sea bed, i.e. dense vegetation, rocks and slopes, needs evaluation. Figure C3.2.42 shows a schematic of a lidar return, as affected by these phenomena and requiring post-processing and evaluation for extraction of water depth to the sea bed. An important consideration is the nadir angle of the scan, which would ideally be fixed to give constant slant range and angle. However, a full conical scan at fixed nadir provides inefficient coverage with more and less dense regions. Various scan patterns have been developed; a modified semi-arc scan, part conical, ahead of the helicopter gives more uniform coverage. Typically 100–200 m swatch width is used with angle from nadir ~20°. For complete evaluation of the sea environment, lidar bathymetry provides a complementary technique to be combined with acoustic and direct mechanical methods of sampling.

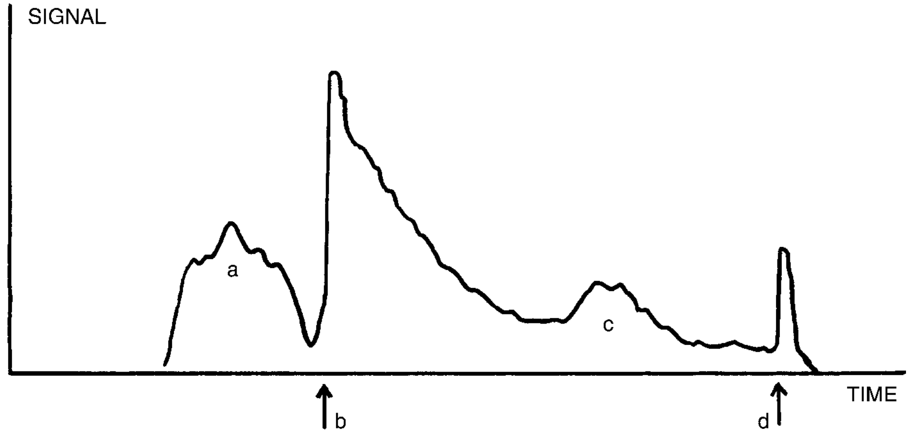


Figure C3.2.42. Schematic of water and seabed scattered return signals: (*a*) haze or “sea smoke”; (*b*) surface return; (*c*) scattering layer; (*d*) seabed return.

C3.2.8.2 Satellite and lunar ranging

The evolution, technology and utility of long-distance, high-precision (near millimetre) ranging to retro-reflectors mounted on satellites (and the moon) has recently been well reviewed by Wood and Appleby [182]. The present brief outline draws heavily on their report. The concept of satellite laser ranging (SLR) to improve geodetic information was first suggested by Plotkin [183]. It was considered that accuracies at the level of a few centimetres would be required to observe interesting geophysical processes of the Earth's crustal dynamics, etc. The first laser radar observations in the mid-1980s demonstrated metre-level precision and established the potential. Over the succeeding years, rapid developments of short-pulse lasers, sensitive detectors, dedicated satellites and accurate computer-guided tracking now give precision approaching a few millimetres, with more than 30 active SLR systems in a worldwide network.

For sub-centimetre precision, very short pulses are required. These are now provided by mode locking systems applied to Nd-YAG lasers giving a train of very narrow (typically ~30 ps) pulses, all equally spaced in an envelope several hundred nanoseconds long. Frequency doubling is employed to take advantage of high quantum efficiency detectors at 0.53 μm in the green. The envelope is approximately Gaussian shaped; in practice, either the most energetic pulse alone may be transmitted or this together with all the following ones. In the latter case, range ambiguity (attribution of which signal return to which pulse) is not a problem since the pulses are separated spatially by 2–3 m, whereas the range accuracy is better than 1 cm. Highly developed techniques are required for the elements of an SLR system and include:

- *Transmitter and receiver telescopes*: they have been coaxial, separate co-mounted and, most recently, separate, individually mounted and controlled [184]. Very high precision, computer-directed acquisition and scanning to follow the satellite path are required.
- *Detectors*: very precise timing of the incoming signal is required. Solid-state devices offer significant improvements over the photo-multipliers used in earlier systems. Single photon detection with low-energy systems is increasingly favoured.
- *Filters*: very narrow, oven-controlled, spectral filters, typically 0.1–0.2 nm wide are employed to cut down sunlight for daytime operation.
- *Retroreflectors*: due to satellite motion, the onboard corner cube retroreflectors require to be slightly 'defocused'. In order to ensure that the footprint of the reflected beam will cover the emitter/receiver station, one of the three mutually perpendicular surfaces is given a slight curvature. Ideally, only one retroreflector would be mounted on the satellite but this gives problems of visibility from different directions. The ultimate solution noted by Wood and Appleby [182] would be eight corner cubes joined at their apices to give reflections equivalent to a reflection from the centre of mass of the satellites.
- *Satellites*: the six dedicated geodetic satellites are typically small (~ 1 m diameter) and spherical, and placed in near-circular orbits at altitudes between 800 and 19 000 km. They are variously suited to different scientific tasks—low orbit: high-frequency gravity terms, ocean tides, etc; higher orbit: crustal motion, Earth rotation, etc.

Several studies of accuracy and range corrections have been given (see, e.g. Refs. [175, 185]). The handbook article of Wood and Appleby [182] provides over 120 references to research articles in satellite laser ranging. One remarkable example was for the NASA LRS system based at Arequipa in Peru. This station was subject to a steady movement of the underlying Nazca plate causing an ENE motion measured at about 10 mm per year. After the earthquake of June 2001, analysis of the dramatic shift in the data confirmed that the station had moved ~0.5 m to the south west.

As to lunar ranging, a number of small arrays of retroreflectors were placed on the moon in the 1960s and early 1970s. Four are in regular use with ranging from two trekking stations—at Mcdonald, Texas and Grasse in France. With the R^{-4} range dependence of signal, ranging to the Moon is vastly more challenging and large telescopes (0.75 and 1.5 m) and high-power pulses (~ 150 mJ) are required.

Several scientific tasks of astrometry can be undertaken, including accurate determination of the Earth and Moon via the tides. The Earth–Moon distance is increasing by 3.8 cm per year and the Earth is slowing on its axis by an increase in the length of day of about 2 ms per century.

C3.2.9 Lidar in space

C3.2.9.1 Introduction—why lidar in space?

During the past 30 years, many passive imagers and sounders have been placed in orbit for study of the Earth's surface and atmosphere with capabilities extending across the visible, IR and microwave spectral regions. Active sensing with radar systems has also been undertaken, with very high-performance scatterometer and imaging radars carried on operational satellites such as ERS-2. Compared with ground-based or even airborne sensors of all types, the prime advantage of space lidar is the prospect of rapid global coverage, including data from otherwise totally inaccessible regions. Compared with other spaceborne sensors (both passive and active), lidar techniques provide certain unique measurement capabilities, such as cloud-top and boundary layer height and global wind fields. Other lidar capabilities (e.g. pressure, temperature and humidity and ranging/altimetry) have the potential to complement and extend existing methods and often provide improved height and spatial resolution, particularly compared with passive techniques.

Against this background, several national and international space agencies have initiated feasibility studies for lidars in space, culminating in several reports, including: LASA (Lidar Atmospheric Sounder and Altimeter), Earth Observing System Vol IId, NASA 1987; LAWS (Laser Atmospheric Wind Sounder), Earth Observing System Vol IIg, NASA 1987; BEST (Bilan Energétique du Système Tropical), CNES, France 1988; Laser Sounding from Space—Report of the ESA Technology Working Group on Space Laser Sounding and Ranging, ESA 1989. This latter report focused on four lidar systems considered good candidates for space deployment; an extract from its executive summary is shown in table C3.2.13. During the past 15 years, several evaluations of the detailed technology and logistics of space lidars for specific tasks have been funded by the space agencies. In addition, various of the critical subcomponents, such as lasers, optical arrangements, lag-angle correction, etc, have been studied and space-qualifiable prototypes constructed. Progress to actual space deployment, however,

Table C3.2.13. The four candidate lidar systems for space considered by ESA in 1989.

(a)	A simple backscatter lidar—for measurements of cloud-top height, cloud extent and optical properties, planetary boundary layer and tropopause height, aerosol distribution—with wide applications to meteorology and climatology
(b)	A differential absorption lidar (DIAL), providing high-vertical-resolution measurements of humidity, temperature and pressure
(c)	A wind-profiling lidar with the unique capability of improved weather forecasting and global dynamics
(d)	A ranging and altimeter lidar for very accurate measurements of surface features, including ground, sea and ice cap height for solid-earth studies

remains slow, with the exception of the NASA Lidar In-space Technology Experiment (LITE), which flew on the Space Shuttle Discovery in September 1994, and very recently the Geoscience Laser Altimetry System (GLAS) launched in 2003.

This and other steps towards spaceborne lidars are discussed in the following sections. Critical design factors for a space lidar obviously include the demands on spacecraft accommodation for mass, size, power, heat dissipation, cooling, pointing and vibrational stability, etc, all for operation in a realistic orbit. Performance factors include ultimate sensitivity limit to weak levels of scattering (largely determined by laser power and telescope aperture) and the geometric terms of range resolution (influenced by laser-pulse length), beam footprint size (beam divergence and the need to keep light levels within eye safe limits—including observation by ground telescopes) and the separation of successive footprints (laser-pulse repetition frequency, beam scanning and spacecraft velocity).

C3.2.9.2 Backscatter lidar and LITE

The science objectives and technology requirements of a spaceborne, direct-detection, backscatter lidar were discussed in the LASA report listed above and also in a 1990 ESA report of the ATLID (Atmospheric Lidar) Consultancy Group. The latter in particular reviewed the potential for operational meteorology and environmental/climatological research. The atmospheric features that should be detectable from such an ATLID backscatter lidar are illustrated in figure C3.2.43 drawn from the report. For weather forecasting (numerical weather prediction (NWP)), ATLID data should be of direct help with clouds, boundary layer height, air mass discontinuities and frontal/convective activity. For climate studies, the information on clouds and aerosols and their effect on the Earth's radiation balance should be a valuable contribution.

As planned (see, e.g. Refs. [186, 187]), the LITE provided a valuable demonstration of the enabling technologies for such operational systems. In the 12-day shuttle mission, the LITE instrument performance was excellent and a large volume of data was generated [72]. Figure C3.2.44 shows a block

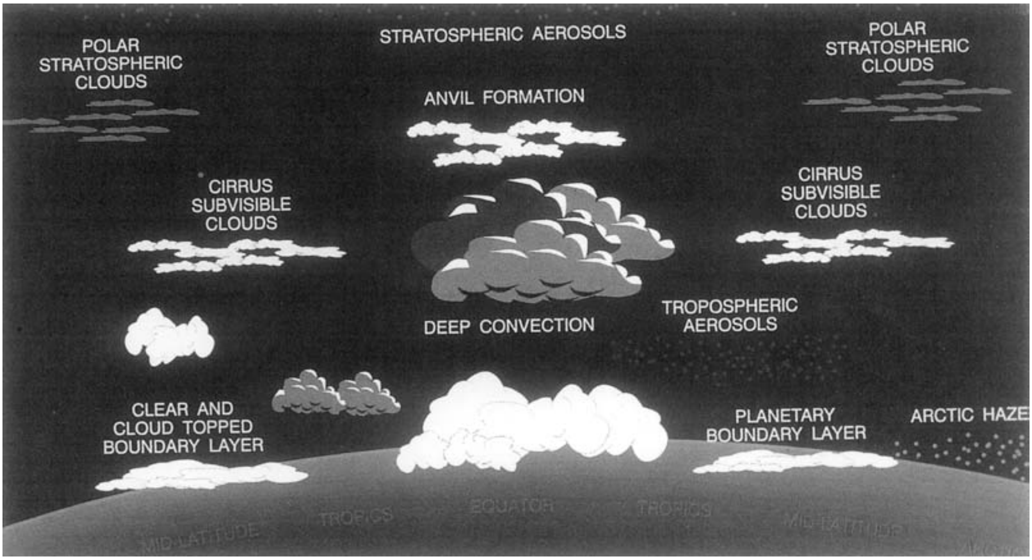

POLAR STRATOSPHERIC CLOUDS
STRATOSPHERIC AEROSOLS
POLAR STRATOSPHERIC CLOUDS
ANVIL FORMATION
CIRRUS SUBVISIBLE CLOUDS
CIRRUS SUBVISIBLE CLOUDS
DEEP CONVECTION
TROPOSPHERIC AEROSOLS
CLEAR AND CLOUD TOPPED BOUNDARY LAYER
PLANETARY BOUNDARY LAYER
ARCTIC HAZE
MID-LATITUDE
TROPICS
EQUATOR
TROPICS
MID-LATITUDE

Figure C3.2.43. Atmospheric features of clouds and aerosols detectable by a simple backscatter lidar at various latitudes (from Report of the 'ATLID' Consultancy Group [195]).

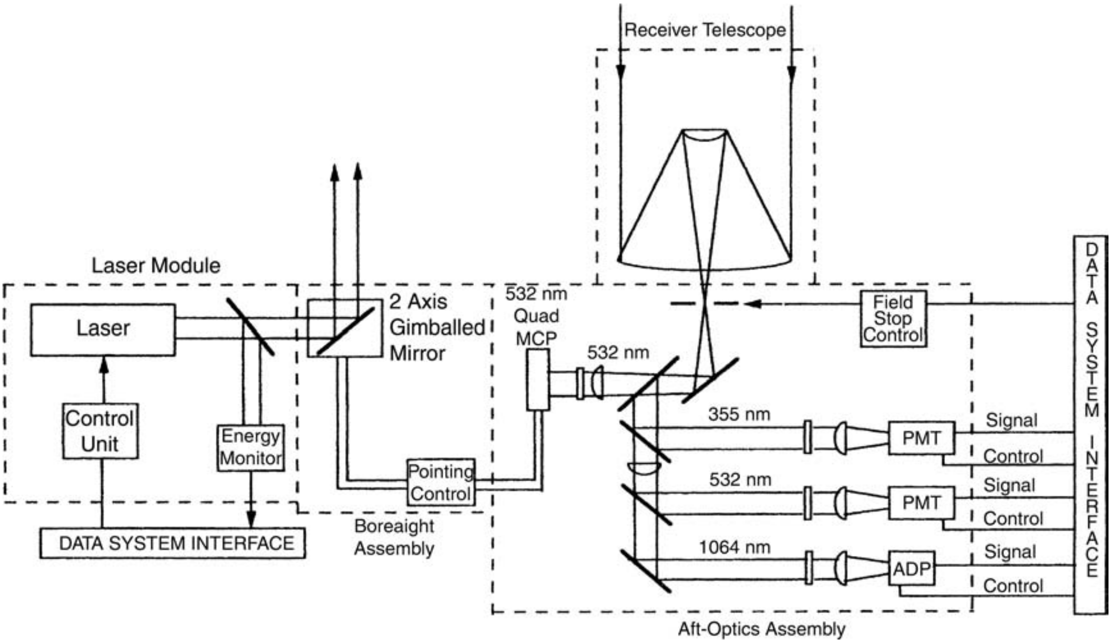

Receiver Telescope
Laser Module
Laser
Control Unit
Energy Monitor
2 Axis Gimballed Mirror
532 nm Quad MCP
532 nm
Field Stop Control
355 nm
532 nm
1064 nm
PMT
PMT
ADP
Signal
Control
Signal
Control
Signal
Control
DATA SYSTEM INTERFACE
Pointing Control
Boreaight Assembly
DATA SYSTEM INTERFACE
Aft-Optics Assembly

Figure C3.2.44. Functional diagram of the LITE system. Light at the three wavelengths was directed down to the Earth from the two-axis gimballed mirror. Scattered light was collected at the lightweight receiver telescope of 1 m aperture and passed via dichroic beam splitters to the two photomultipliers (PMT) and avalanche photodiode (ADP). Some of the 0.532 μm light was passed to the microchannel plate (MCP) quadrant detector, which provided an error signal for precise alignment of the transmitter and receiver system (from Ref. [72]).

diagram of the equipment, which was mounted on a Spacelab pallet. The flash-lamp-pumped Nd-YAG laser operated in doubled and tripled frequency mode to give simultaneous output pulses at 1.064, 0.532 and 0.355 μm. The pulse repetition rate was 10 s^{-1} and pulse width was 27 ns. The beam footprint at the Earth's surface varied between 290 m (0.355 μm) and 470 m (1.06 μm) and successive footprints were 740 m apart. The mass of the instrument was 990 kg and the average power consumption in lasing operation was 3.1 kW.

Two of the fascinating records from this instrument are shown in figure C3.2.45 and figure C3.2.46. These pictures provide some indication of the quality, depth and volume of information that could be available from an operational system, and indeed the extent of the signal processing, evaluation and transmission systems that would need to be set up to use it effectively. For meteorological and weather forecasting purposes, such information must of course be made available in good time to the operational centres.

C3.2.9.3 Global wind field measurement by lidar

One of the earliest feasibility studies of a satellite-borne lidar for global wind measurement was sponsored by the USAF Defence Meteorological Satellite Program. In the early 1980s, this was followed by a hardware definition study, conducted by Lockheed, of a system given the acronym WINDSAT (Windmeasuring Satellite); another title current at this time was WPLID for Wind Profiling Lidar (see, e.g. Ref. [103]). These have been followed by the reports noted in C3.2.9.1, all of which considered large, multi-scanning (conical or several axes) lidars. More recently, the ESA Atmospheric Dynamics

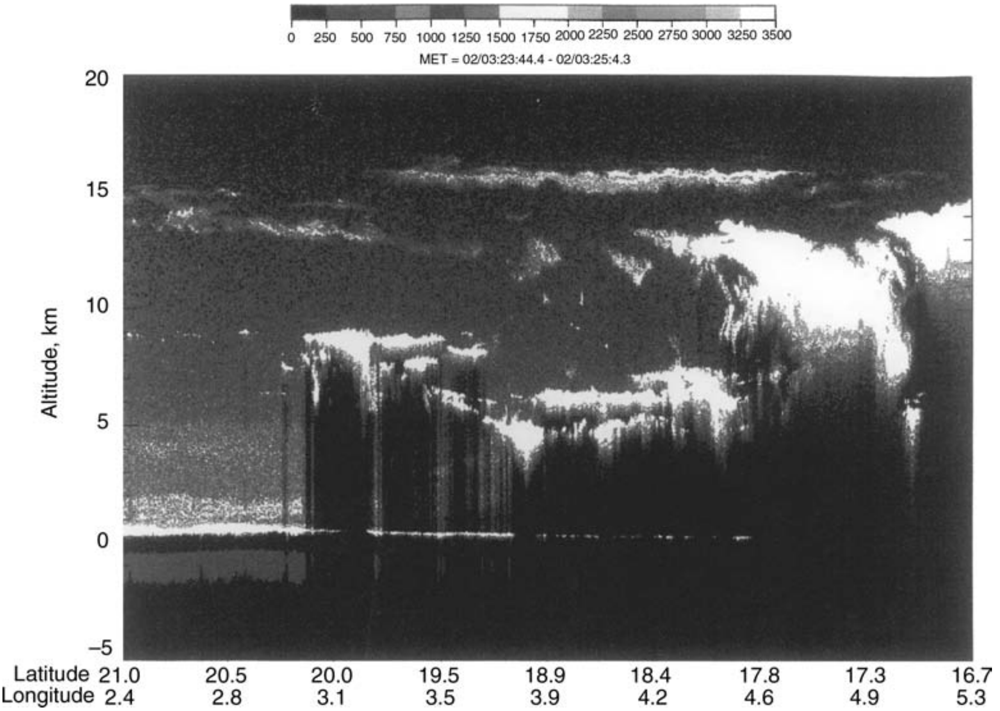


Figure C3.2.45. The LITE return signal at 0.532 μm after offset correction showing the multilayered cloud structure associated with a tropical storm system over West Africa. Note the aerosol signal (yellow and red) on the left-hand side, particularly at the boundary layer. Ground signals are also obtained beneath the high-level thin cloud (16 km, at centre of picture), but not below the thicker cloud layers on the right hand side (from Ref. [72]).

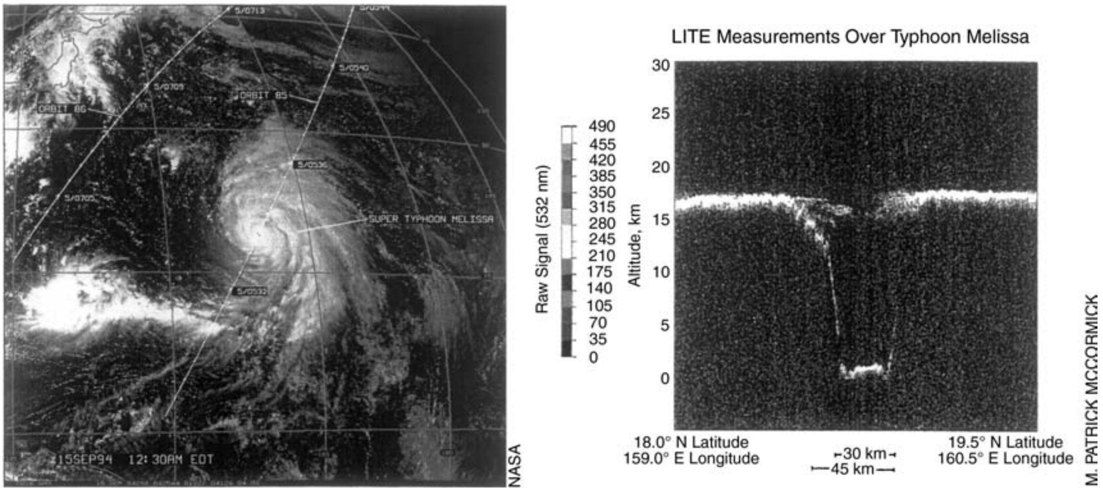


Figure C3.2.46. The LITE system serendipitously passed directly over the eye of typhoon Melissa. Satellite photographs were obscured by thin cloud. Nevertheless, the lidar was able to penetrate this and make measurements within the eye-wall down to the surface (shown by Barnes and Hofmann [31]).

Mission: Reports for Mission Selection (1999) was presented to the European Earth Observation community at a meeting attended by over 300 participants. Of the four candidate missions, ADM was selected as the second Core Mission to be implemented. The lidar studies indicate the degree of importance attached to global scale wind measurement deriving from the unique capabilities of a Doppler Wind Lidar (DWL) compared with other spaceborne methods. For example, low-level wind strength and direction may be derived from radar scattering from the sea surface where the wave height is largely determined by the surface wind. However, such low-level winds are not of the highest value for meteorological purposes. In the mid-levels, the winds derived from cloud-top motion are not always completely representative of the air motion in that region, and may also have some uncertainty as to relevant altitude. For the ESA ADM, the observational requirements have been chosen as shown in table C3.2.14 and led to selection of a relatively simple single-axis system for the first demonstrator mission. The most suitable lidar to realize these goals has been extensively discussed with three possible candidates:

(a) Coherent heterodyne, with CO_2 gas lasers in the 10 um band (aerosol scattering).

(b) Coherent heterodyne, with solid-state slab lasers in the 1–2 um band (aerosol scattering).

(c) Incoherent direct detection, at ~0.3–0.5 um, and Fabry–Perot interferometer/edge technique analysis (aerosol and molecular scattering).

For the coherent systems, the laser beam must be transmitted in a collimated, single-mode beam. From a 400–500 km orbit, the beam size, or footprint, at the Earth's surface would be of order 10 m, which rules out a non-eye-safe laser of wavelength less than ~1.5 μm. For the incoherent systems at shorter wavelength, the beam may be expanded to produce an enlarged footprint with energy density within eye-safe limits. For any system, the crucial questions derive from the backscattering characteristics of the atmosphere that determine the primary Doppler signal, whether from aerosols or molecules.

As discussed in section C3.2.4, molecular scattering varies uniformly and predictably whereas aerosol scattering is highly variable with occasional large increases due to volcanic activity (a suggestion

Table C3.2.14. Observational requirements of the ESA atmospheric dynamics mission.

		Observational Requirements		
		PBL	Troposphere	Stratosphere
Vertical domain	km	0–2	2–16	16–20
Vertical resolution	km	0.5	1.0	2.0
Horizontal domain			Global	
Number of profiles	h^{-1}		>100	
Profile separation	km		>200	
Horizontal integration length	km		50	
Horizontal sub-sample length	km		0.7–50	
Accuracy (HLOS component)	$m\ s^{-1}$	1	2	3
Zero-wind bias	$m\ s^{-1}$		0.1	
Windspeed slope error	%		0.5	
Data reliability	%		95	
Data availability	h		3	
Length of observation data set	yr		3	

ALADIN : LAUNCH - 5 DAYS...
REALLY, MICHAEL, YOU THINK IT'S ENOUGH TO ENHANCE β
OH SURE, JEAN-YVES, WE OUGHT TO GET AN ORDER OF MAGNITUDE AT LEAST !
PINATUBO
ALADIN
GSE

Figure C3.2.47. A suggestion for augmenting aerosol scattering β in the atmosphere to improve the performance of the spaceborne Doppler wind lidar (ALADIN). Courtesy of Rodolphe Krawczyk, 1994.

for augmenting the scattering levels with a strategically placed device is shown in figure C3.2.47). For both domains, a large, powerful lidar is required and may be quantified in terms of the necessary ED^2, where E is the total laser pulse energy (Joules per individual measurement) and D is the collecting telescope diameter (metres). For successful measurement at lowest levels of atmospheric scattering (both molecular and aerosol), analysis shows that ED^2 requires to be greater than $\sim 100\,\mathrm{J\,m^2}$. With $D^2 \sim 1$, for the heterodyne systems the available energy is best utilized in large pulses, ideally > 10 J, whereas for direct detection lidars the energy may be distributed over many small pulses (see section C3.2.3.4).

Extensive modelling of performance for spaceborne wind lidars has been conducted. Modelled performance for different levels of aerosol scattering—median, quartiles, deciles—corresponding to the values shown in figure C3.2.8 and for a coherent heterodyne lidar operating with the instrumental parameters listed is shown in figure C3.2.48.

Most recently, for the ADM demonstrator, ESA has chosen a direct-detection interferometric lidar operating at 0.35 μm; the basic optical layout is shown in figure C3.2.49. In a novel arrangement, the narrow band aerosol (Mie) scattering would be separately analysed by a Fizeau or possibly a two-beam Michelson interferometer; the broadband molecular (Rayleigh) scattering, which is relatively stronger at these shorter wavelengths, would be analysed by a double-edge Fabry–Perot etalon. Suggested performance is shown in figure C3.2.50 with an accuracy better than $2\,\mathrm{m\,s^{-1}}$ up to 11 km altitude. An artist's impression of this equipment in a satellite flying in a sun-synchronous polar orbit at an altitude of ~400 km is shown in figure C3.2.51. Processing of the backscatter signals will provide about 3000 globally distributed wind profiles per day, above thick cloud or to the surface in clear air, at typically 200 km separation along the satellite track. Target date for launch is 2008.

Parameters used for the performance assessment

OrbAlt = 400°km	***Orbit altitude***
θ = 30°deg	***Nadir angle***
λ = 9.1°µm	***Laser wavelength***
E = 22.5°joule	***Pulse energy***
n = 5	***Number of pulses accumulated over 25 km horizontally***
D = 0.8°m	***Telescope diameter***
w = 0.35°MHz	***Return signal bandwidth (HW@ exp(–0.5))***
ηl = 0.12	***Overall instrument efficiency (optics transmission, overall heterodyne efficiency)***
Bn = 40°MHz	***Processing bandwidth***
ΔH = 1°km	***Vertical resolution***

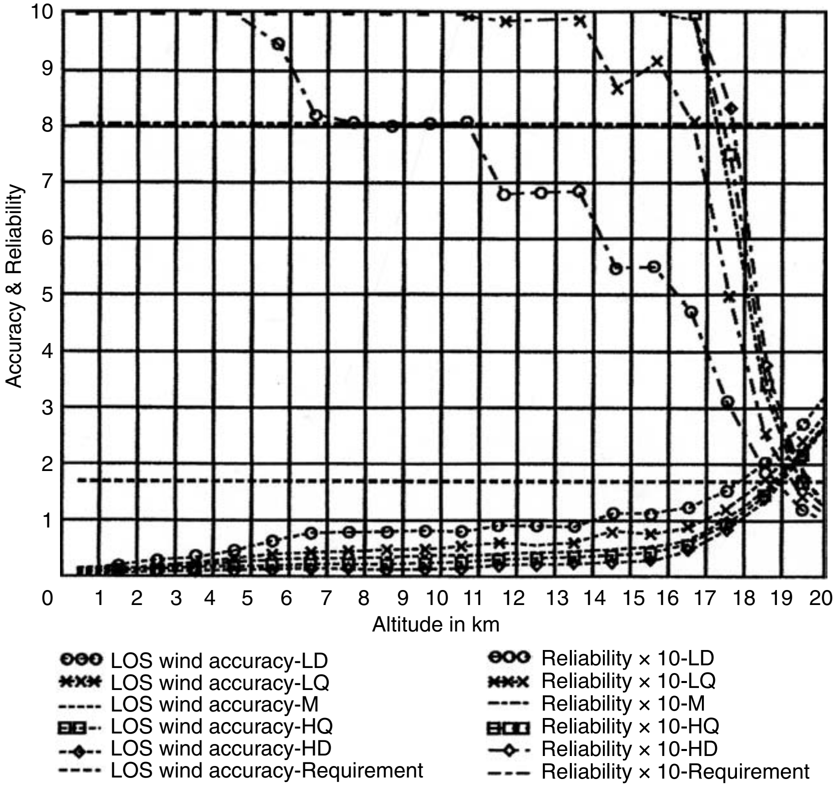


Figure C3.2.48. Calculated performance of a space-borne 9.1 µm coherent detection lidar with parameters shown for median and upper/lower quartiles and deciles of aerosol scattering. Note that the accuracy (ms^{-1}) curves start at the lower left-hand side and the reliability curves (0–100%) start at the top left-hand side for the different levels of scattering.

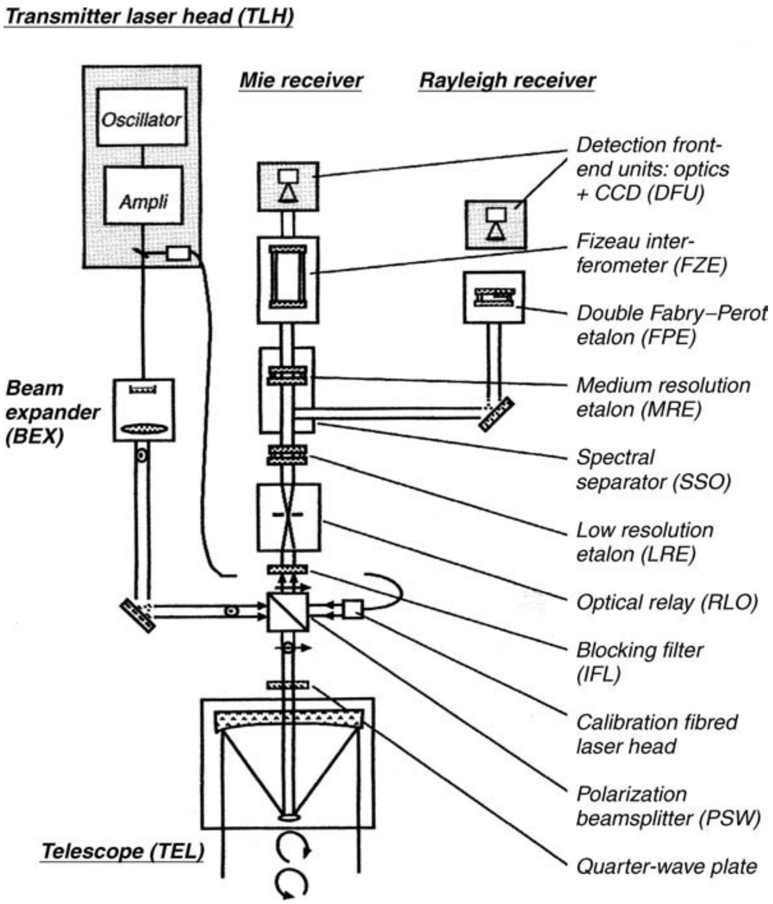

Transmitter laser head (TLH)
Oscillator
Ampli
Mie receiver
Rayleigh receiver
Detection front-end units: optics + CCD (DFU)
Fizeau interferometer (FZE)
Double Fabry–Perot etalon (FPE)
Beam expander (BEX)
Medium resolution etalon (MRE)
Spectral separator (SSO)
Low resolution etalon (LRE)
Optical relay (RLO)
Blocking filter (IFL)
Calibration fibred laser head
Polarization beamsplitter (PSW)
Telescope (TEL)
Quarter-wave plate

Figure C3.2.49. Optical arrangement for the ALADIN direct detection lidar for space-borne Doppler wind lidar. The aerosol (Mie) and molecular (Rayleigh) scattered signal are measured in the two separate interferometer systems as shown.

C3.2.9.4 Chemical, ranging and solid earth studies

As noted in table C3.2.13, DIAL lidar is an obvious candidate for measurements into the atmosphere from space (see, e.g. Ref. [188]). The problems, however, are formidable: two or more laser wavelengths for each chemical species under study (ozone, water vapour, greenhouse gases, etc), compensations for the Doppler shifts in the beams due to the large spacecraft velocity (typically 15° off nadir $\equiv 0.3\,\text{cm}^{-1}$ shift), with the beams transmitted simultaneously and co-axially to ensure return signals from the same elastic scatterers. Much study has been put into such active remote sensing from space and DIAL lidars will almost certainly be flown in the coming years. However, passive techniques over a wide range of wavelengths are providing much information. For example, in the NASA Earth Observing System (EOS), the Aura mission to study chemistry and dynamics of the troposphere and stratosphere, the instruments include sounders for measurement of thermal emission from the atmospheric limbs (microwave and high resolution) as well as nadir viewing imaging spectrographs for ozone and trace gases.

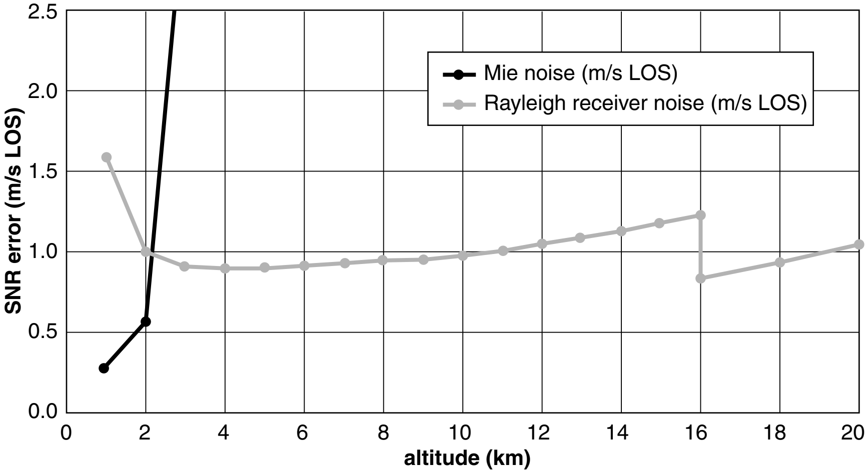


Figure C3.2.50. Calculated performance, by Monte Carlo simulation for the ALADIN space-borne Doppler wind lidar. For this calculation, the main parameters were: a 1.1 m diameter telescope, over 700 shots a total pulse energy per measurement of 120 J at 0.35 μm wavelength and a detector quantum efficiency of 80%. Performance may be readily scaled by the Cramer–Rao relationship. The shift at 16 km is due to a change in the required vertical resolution from 1 to 2 km at that altitude in the atmosphere [196].

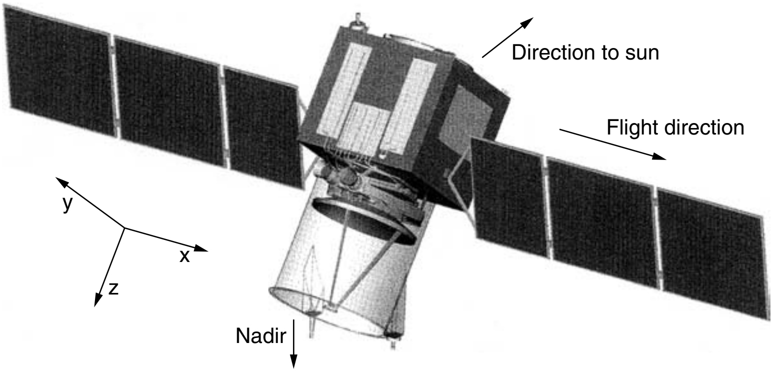


Figure C3.2.51. Impression of the in-orbit configuration for a space-borne Doppler wind lidar showing the telescope pointing down at ~35° from nadir [196].

Lidar for geoscience has been outlined in section C3.2.8. Early studies for space-borne systems were given by Cohen *et al* [189] and Bufton [175]. Very recently, in February 2003, the Geoscience Laser Altimetry System (GLAS) on board the NASA Ice, Cloud and Land Elevation Satellite (ICE Sat) was activated. This mission is intended to accurately measure ice sheet mass balance, cloud and aerosol heights as well as land topography and vegetation characteristics. Initial test of the space-bound lidar altimeter system was validated with data from the global Laser Reference System (LRS) noted in section C3.2.8. The lidar operates on Nd-YAG at 1.06 μm; it is reported that the signal returns were as expected and the transmitted laser beam is very close to the bore sight.

C3.2.10 Conclusions

From the preceding papers cited, it should be clear that the past 30 years have seen remarkable advances in laser radar technology, applications and understanding. Much fine equipment has been installed in different centres worldwide and valuable, often unique, measurements not accessible by other techniques have been made. However, there have also been significant disappointments, particularly in the wider deployment of coherent lidar equipments and techniques. This final section attempts a brief analysis of the situation and to draw some guidelines for the future.

Table C3.2.15 offers a crude summary comparison of the history and development of four remote-sensing technologies. It is striking that, for microwave radar, thermal imaging and sodar, the first deployment (defined as use by non-specialist scientists and technicians) took place within ~15 years, followed by widespread deployment (e.g. equipments commercially available off-shelf or bespoke) less than 10 years later. Examination of a radar supplier's catalogue with the dozens of different equipments, all available off-the-shelf, for many varied applications working over a large range of wavelengths provides a salutary example to the lidar scientist. Even closer is the astonishing growth in the optical technology and applications for the telecommunications industry. However, for laser radar, apart from relatively simple tasks of military and civil ranging, target marking and possibly DIAL, widespread deployment has scarcely been achieved. The question is why should this be so? Are more advanced lidar systems perceived as too difficult, too expensive, unreliable, dangerous, too uncertain, etc?

As scientists, we tend to believe the adage 'build a better mousetrap and the world will beat a path to your door'. In the hard, commercial world, this is not true—a new technology has to be significantly superior to displace an older established one and is unlikely to do so if it is even marginally more

Table C3.2.15. Comparison of remote sensing technologies.

Technology	First Research	First Deployment (a)	Widespread Deployment (b)	Comment
Microwave radar	~ 1935 'kill a sheep at 2 miles'	~ 1940—military 'needs a Ph.D. in every set'	>1942—aided by compact cavity magnetrons	Enormous variety of capabilities, application wavelengths, equipment, etc.
Thermal imaging	1930–1970	~1975	~1985	Widespread applications in civil and military domains.
Active acoustic sensors (SODAR)	1960–1970	~1975	~1985	Slow, limited range, low resolution etc *BUT* relatively cheap and simple.
Direct detection lidar: (a) temporal	1965–1975	~1975	~1980	Military and civil ranging, target marking, cloud base, etc.
(b) spectral	1970–1985	~1985	———	Chemical detection (DIAL); limited Doppler applications.
Coherent lidar	1970–1990	~2000	———	Enormous range of applications demonstrated. Why not taken up outside research?

(a) Defined as first use by non-specialist scientists/technicians. (b) Defined as widely available off-shelf (or bespoke).

expensive. Other relevant truisms that bear heavily on the planning, specification and design of equipments include:

- The 'best' is the enemy of the 'good'.
- The devil is in the detail.
- Small is beautiful – but is it achievable?
- Multifunction usually means, literally, good for nothing.

Ignoring one or more of these is usually a recipe for disaster — as is very evident in a review conducted by the author (unpublished and probably unpublishable!) of nine 'failed' projects. These were all large, expensive (equivalent of several million pound sterling or more) projects funded during the past 25 years by civil and military agencies across Europe and the United States; all were intended to lead into wider development and even (in some cases) medium-scale production. The technical performances of the prototypes in these projects ranged from poor to very good indeed. For those in the latter category, lack of ongoing success was basically due in one case to failure to sufficiently interest the end user (partly due to the inadequate data/display facilities available at the time) and, for the other, an alternative microwave radar technology was preferred. In at least two projects, the basic properties of the phenomenon under investigation were poorly researched and understood (scattering strength, spectral form, etc) so the lidars, when built to the set specification, could not provide the required information. In two others, the performance specification became overly ambitious (the 'best') and multifunction leading to excessive size, complexity and cost, and poor overall performance. Other projects were bedevilled by arguments between lidar scientists over the appropriate choice of laser, not helped by inflated and premature claims from laser theorists and manufacturers as to the achievable pulse energy, prf, reliability, etc. In one bad example, basic principles of coherent operation were not properly appreciated at a high technical level, leading to poor decision making. From this brief history, a few more truisms may be drawn:

- A project will fail with too much technology 'push' and not enough end user 'pull'.
- Bad science and over-promoted technology will breed disaster—but good science and technology will not necessarily win.
- Beware 'technology jocks' bearing gifts.
- The commendation 'more reliable' must mean 'an in-depth engineering study of failure modes has been conducted' or it is worse than useless.

One may thus draw a few guidelines (see section C3.2.5 in the present volume for a much more comprehensive analysis): the good system and project development arises from a total holistic approach incorporating:

(1) A good understanding of the measurement problem with a realistic specification for the lidar based on reliable, available component technology.

(2) Good science/technology in the lidar design including:
    - Good fundamental science of direct detection/coherent lidar.
    - Translation into simple, serviceable optical design.

- Proper study of engineering failure modes.
- Good signal processing.
- Good data analysis and display which provide:
  - Swift timely analysis.
  - Informative graphics.
  - And truly give the end user what he needs and can use.

This, of course, raises various broader questions including, for example:

- Where is the basic lidar training? Few universities conduct research with coherent lidar, which seems a little puzzling, particularly with the potential for laboratory-based physics applications of coherent heterodyne techniques.
- Is there an adequate, compact and referable body of knowledge? An enormous body of research publications and reviews exists, but a more pedagogic literature is perhaps lacking.
- How to involve and convince end-users? The bulk of lidar applications research is still done by lidar scientists who initially built demonstrators but then developed an interest in, e.g. the applications to meteorology, atmospheric chemistry or wake vortices, etc, occasionally managing to interest colleagues more directly involved in these fields.
- This final point emphasizes that, without an enthusiastic body of such end-users (professional meteorologists, environmental scientists, system engineers, etc), who can envisage worthwhile application of a lidar to 'their' problem, appreciable funding is not going to be forthcoming from the research councils and agencies or from industry.

Lest this be considered too downbeat, an assessment of lidar must emphasize again some of the remarkable and unique achievements of laser radar over the past 30 years: contributions to flow measurement, wind fields, atmospheric chemistry, ozone depletion, pollution dispersal, avionics, geoscience, etc, even apart from more military-oriented applications. Looking to the future, progress is likely to be rooted in:

- Laser developments: particularly of high power and efficient, agile-frequency CO_2 lasers, precisely specified off-the-shelf solid-state 2 and 1 μm lasers, and diode and fibre amplifier lasers at $\sim$1.5 μm with increased power and mode purity.
- Lidar construction: reduction in cost and complexity should flow from monolithic waveguide-type construction at 10 μm and, at shorter wavelengths, much use of fibre guiding, components and optical techniques borrowed from the communications industry.
- Detectors: very fast, high quantum efficiency ($>$80%), low-noise array detectors are becoming available at visible and UV wavelengths. Extension to the near-IR would have great impact, as would uncooled heterodyne detectors at 10 μm.
- Data analysis and display: the rapid, user-friendly access to data coupled with graphic display of results now offered by modern computing facilities provides a great advance for lidar studies and should always be budgeted for in a system programme.

Finally, it is worth noting the encouraging prospects for global wind field measurements in the Atmospheric Dynamics Mission – ESA's technology demonstrator, due to be launched in 2008, and at the other extreme the potential development of small, low-cost LDVs for flow measurement ahead of, for example, wind-turbine generators.

Acknowledgments

The author expresses his deep appreciation of the constant spur of enthusiasm and interest generated by many colleagues and friends worldwide, with particular thanks to those at RSRE/DERA (Malvern)—now QINETIQ—and at ESTEC (Noordwijk). As always, any errors or omissions in the present work must be attributed to the author.

References

[1] Hermann J A 1990 Evaluation of the scintillation factor for laser hazard analysis *Appl. Opt.* **29** 1287–1292
[2] Rye B J and Hardesty R M 1997 Estimate optimisation parameters for incoherent backscatter heterodyne lidar *Appl. Opt.* **36** 36
[3] Vaughan J M 2002 Scattering in the atmosphere *Scattering* Chapter 2.4.3 eds E R Pike and P Sabatier (London: Academic Press) pp 937–957
[4] Harris M 1995 Light-field fluctuations in space and time *Contemp. Phys.* **3** 215–233
[5] Schwiesow R L and Cupp R E 1980 Calibration of a cw infrared Doppler lidar *Appl. Opt.* **19** 3168–3172
[6] Hardesty R M, Keeler R J, Post M J and Richter R A 1981 Characteristics of coherent lidar returns from calibration targets and aerosols *Appl. Opt.* **20** 3763–3769
[7] Kavaya M J and Menzies R T 1985 Lidar aerosol backscatter measurements: systematic, modelling and calibration error considerations *Appl. Opt.* **26** 796–804
[8] Vaughan J M, Callan R D, Bowdle D A and Rothermel J 1989 Spectral analysis, digital integration, and measurement of low backscatter in coherent laser radar *Appl. Opt.* 15 **28** 3008–3014
[9] Sonnenschein C M and Horrigan F A 1971 Signal-to-noise relationships for coaxial systems that heterodyne backscatter from the atmosphere *Appl. Opt.* **10** 1600–1604
[10] Kavaya M J and Suni P J M 1991 Continuous wave coherent laser radar: calculation of measurement location and volume *Appl. Opt.* **30** 2634–2642
[11] Werner C, Kopp F and Schwiesow R L 1984 Influence of clouds and fog on LDA wind measurements *Appl. Opt.* **23** 2482–2484
[12] Chan K P and Killinger D K 1992 Useful receiver telescope diameter of ground-based and airborne 1-, 2-, and 10-μm coherent lidars in the presence of atmospheric refractive turbulence *Appl. Opt.* **31** 4915–4917
[13] Frehlich R and Kavaya M J 1991 Coherent laser radar performance for general atmospheric refractive turbulence *Appl. Opt.* 36 **30** 5325–5352
[14] Scorer R S 1997 Dynamics of meteorology and climate *Atmospheric Physics* Wiley-Praxis Series(Chichester: Wiley)
[15] Hinkley E D ed 1976 *Laser Monitoring of the Atmosphere* (Berlin: Springer)
[16] Zuev V E 1976 Laser light transmission through the atmosphere *Laser Monitoring in the Atmosphere* ed E D Hinkley (Berlin: Springer)
[17] Collis R T H and Russell P B 1976 Lidar measurements of particles and gases by elastic backscattering and differential absorption *Laser Monitoring in the Atmosphere* ed E D Hinkley (Berlin: Springer)
[18] Inaba H 1976 Detection of atoms and molecules by Raman scattering and resonance fluorescence *Laser Monitoring in the Atmosphere* ed E D Hinkley (Berlin: Springer)
[19] Fabelinski I L 1968 *Molecular Scattering of Light* (New York: Plenum Press)
[20] Rye B J 1998 Molecular backscatter heterodyne lidar: a computational evaluation *Appl. Opt.* **37** 27
[21] Chanin M L, Hauchecorne A, Garnier A and Nedelkovic D 1994 Recent lidar developments to monitor stratosphere-troposphere exchange *J. Atmos. Terrest. Phys.* **56** 1073–1081
[22] Browell E V, Ismail S and Grant W B 1998 Differential absorption lidar (DIAL) measurements from air and space *Appl. Phys.* **67** 399–410
[23] Seinfeld J H and Pandis S N 1998 *Atmospheric Chemistry and Physics: from Air Pollution to Climate Change* (New York: Wiley)
[24] Malcolm A L, Derwent R G and Maryon R H 2000 Modelling the long-range transport of secondary PM10 to the UK *Atmos. Environ.* **34** 881–894

[25] Derwent R G, Simmonds P G, Seuring S and Dimmer C 1998 Observation and interpretation of the seasonal cycles in the surface concentrations of ozone and carbon monoxide at Mace Head Ireland, from 1990 to 1994 *Atmos. Environ.* **31** 145–157

[26] Ryall D B, Maryon R H, Derwent R G and Simmonds P G 1998 Modelling long-range transport of CFCs to Mace Head, Ireland *QJR Meteorol. Soc.* **124** 417–446

[27] Heffter J L and Stunder B J B 1993 Volcanic ash forecast transport and dispersion (VAFTAD) model *Computer Techniques* **8** 533–541

[28] Mosca S, Bianconi R, Bellasio R, Graziani G and Klug W 1998 *ATMES II – Evaluation of long-range dispersion models using data of the first EXTEX release, EC EUR 17756 EN*

[29] Maryon R H and Buckland A T 1995 Tropospheric dispersion: the first ten days after a puff release *QJR Meteorol. Soc.* **121** 1799–1833

[30] Ansmann A, Mattis I, Wandinger U, Wagner F, Reichardt J and Deshler T 1997 Evolution of the Pinatubo aerosol: Raman lidar observations of particle optical depth, effective radius, mass, and surface area over Central Europe at 53. 4N *J. Atmos. Sci.* **54** 2630–2641

[31] Barnes J E and Hofmann D J 1997 Lidar measurements of stratospheric aerosol over Mauna Loa Observatory *Geophys. Res. Lett.* **24** 1923–1926

[32] Browell E V, Fenn M A, Butler C F and Grant W B 1996 Ozone and aerosol distributions and air mass characteristics over the South Atlantic basin during the burning season *J. Geophys. Res.* **101** 24043–24068

[33] Browell E V, Gregory G L, Harris R C and Kirchhoff V W J H 1998 Tropospheric ozone and aerosol distributions across the Amazon Basin *J. Geophys. Res.* **93** 1431–1452

[34] Cutten D R, Pueschel R E, Bowdle D A, Srivastava V, Clarke A D, Rothermel J, Spinhirne J D and Menzies R T 1996 Multiwavelength comparison of modelled and measured remote tropospheric aerosol backscatter over Pacific Ocean *J. Geophys. Res.* **101** 9375–9389

[35] Donovan D P, Bird J C, Whiteway J A, Duck T J, Pal S R, Carswell A I, Sandilands J W and Kaminski J W 1996 Ozone and aerosol observed by lidar in the Canadian Artic during the winter of 1995/96 *Geophys. Res. Lett.* **23** 3317–3320

[36] Hoff R M, Harwood M, Sheppard A, Froude F, Martin J B and Strapp W 1997 Use of airborne lidar to determine aerosol sources and movement in the Lower Fraser Valley (LFB), BC *Atmos. Environ.* **31** 2123–2134

[37] Li S, Strawbridge K B, Leaitch W R and Macdonald A M 1998 Aerosol backscattering determined from chemical and physical properties and lidar observations over the east coast of Canada *Geophys. Res. Lett.* **25** 1653–1656

[38] Marenco F, Santacesaria V, Bais A, Balis D, di Sarra A and Papayannis A 1997 Optical properties of tropospheric aerosols determined by lidar and spectrophotometric measurements (PAUR campaign) *Appl. Opt.* **36** 6875–6886

[39] McCormick M P, Thomason L W and Trepte C R 1995 Atmospheric effects of the Mt Pinatubo eruption *Nature* **373** 399–404

[40] Osborn M T, Kent G S and Trepte C R 1998 Stratospheric aerosol measurements by the lidar in space technology experiment *J. Geophys. Res.* **103** 11447–11454

[41] Parameswaren K, Rajan R, Vijayakumar G, Rajeev K, Moorthy K K, Nair P R and Satheesh S K 1998 Seasonal and long term variations of aerosol content in the atmospheric mixing region at a tropical station on the Arabian sea coast *J. Atmos. Solar-Terrest. Phys.* **60** 17–25

[42] Post M J, Grund C J, Wang D and Deshler T 1997 Evolution of Mount Pinatubo's aerosol size distributions over the continental United States; two wavelength lidar retrievals and in situ measurements *J. Geophys. Res.* **102** 13535–13542

[43] Shibata T, Itabe T, Mizutani K, Uchino O, Nagai T and Fujimoto T 1996 Arctic tropospheric aerosols and clouds in the polar night season observed by a lidar at Eureka, Canada *J. Geomagn. Geoelectr.* **48** 1169–1177

[44] Spinhirne J D, Chudamani S, Cavanaugh J F and Bufton J K 1997 Aerosol and cloud backscatter at 1.06, 1.54 and 0.53μm by airborne hard-target-calibrated Nd: YAG/methane Raman lidar *Appl. Opt.* **36** 3475–3490

[45] Post M J 1984 Aerosol backscattering profiles at CO_2 wavelengths: the NOAA data base *Appl. Opt.* **23** 2507–2509

[46] Post M J 1986 Atmospheric purging of El Chichon debris *Journal of Geophysical Research* **91** 5222–5228

[47] Rothermel J, Bowdle D A, Vaughan J M and Post M J 1989 Evidence of a tropospheric aerosol backscatter background mode *Appl. Opt.* **28** 1040–1042

[48] Bowdle D A, Rothermel J, Vaughan J M and Post M J 1991 Aerosol backscatter measurements at 10.6 micrometers with airborne and ground-based CO_2 Doppler lidars over the Colorado High Plains 2. Backscatter Structure *J. Geophys. Res.* **96** 5337–5344

[49] Bilbro J W, DiMarzio C, Fitzjarrald D, Johnson S and Jones W 1986 Airborne Doppler lidar measurements *Appl. Opt.* **25** 3952–3960

[50] Rothermel J, Bowdle D A and Srivastava V 1996 Midtropospheric aerosol backscatter background mode over the Pacific Ocean at 9.1 μm wavelength *Geophys. Res. Lett.* **23** 281

[51] Srivastava V, Jarzembski M A and Bowdle D A 1992 Comparison of calculated aerosol backscatter at 9.1 and 2.1μm wavelengths *Appl. Opt.* **31** 11904–11906

[52] Srivastava V, Rothermel J, Bowdle D A, Jarzembski N A, Chambers D M and Clarke A D 1995 High resolution remote sensing of sulphate and aerosols form CO_2 lidar backscatter *Geophys. Res. Lett.* **22** 2373–2376

[53] Srivastava V, Rothermel J, Jarzembski M A, Chambers D M and Clarke A D 1997 Comparison of modelled backscatter using measured aerosol microphysics with focused cw lidar data over Pacific *J. Geophys. Res.* **102** 16605–16617

[54] Gras J L, Platt C M, Jones W D, Huffaker R M, Young S A, Banks S M and Booth D J 1991 Southern Hemisphere, tropospheric aerosol backscatter measurements – Implications for a laser wind system *J. Geophys. Res.* **96** 5357–5367

[55] Alejandro S B, Koenig G G, Vaughan J M and Davies P H 1990 SABLE: A South Atlantic aerosol backscatter measurement programme *Bull. Am. Meteorol. Soc.* **71** 281–287

[56] Alejandro S B, Koenig G G, Bedo D, Swirbalus T, Frelin R, Woffinden J, Vaughan J M, Brown D W, Callan R, Davies P H, Foord R, Nash C and Wilson D J 1995 Atlantic atmospheric aerosol studies 1. Programme overview and airborne lidar *J. Geophys. Res.* **100** 1035–1041

[57] Vaughan J M, Brown D W, Nash C, Alejandro S B and Koenig G G 1995 Atlantic atmospheric aerosol studies 2. Compendium of airborne backscatter measurements at 10.6μm *J. Geophys. Res.* **100** 1043–1065

[58] Gibson F W 1994 Variability in atmospheric light-scattering properties with altitude *Appl. Opt.* **23** 411–418

[59] Menzies R T, Tratt D M and Flamant P H 1994 Airborne CO_2 coherent lidar measurements of cloud backscatter and opacity over the ocean surface *J. Atmos.Oceanic Technol.* **11** 770–778

[60] Menzies R T and Tratt D M 1994 Airborne CO_2 coherent lidar for measurements of atmospheric aerosol and cloud backscatter *Appl. Opt.* **33** 5698–5711

[61] Menzies R T and Tratt D M 1995 Evidence of seasonally dependent stratosphere-troposphere exchange and purging of lower stratospheric aerosol from a multi-year lidar data set *J. Geophys. Res.* **100** 3139–3148

[62] Menzies R T and Tratt D M 1997 Airborne lidar observations of tropospheric aerosols during the Global Backscatter Experiment (GLOBE) Pacific circumnavigation missions of 1989 and 1990 *J. Geophys. Res.* **102** 3701–3714

[63] Tratt D M and Menzies R T 1995 Evolution of the Pinatubo volcanic aerosol column above Pasadena, California observed with a mid-infrared backscatter lidar *Geophys. Res. Lett.* **22** 807–881

[64] Vaughan J M, Brown D W, Davies P H, Nash N, Kent G and McCormick M P 1988 Comparison of SAGE II solar extinction data with airborne measurements of atmospheric backscattering in the troposphere and lower stratosphere *Nature* **332** 709–711

[65] Vaughan J M, Steinwall K O, Werner C and Flamant P H 1997 Coherent laser radar in Europe *Proc. IEEE* **84** 205–226

[66] Vaughan J M, Brown D W and Willetts D V 1998 The impact of atmospheric stratification on a space-borne Doppler wind lidar *J. Mod. Opt.* **45** 1583–1599

[67] Vaughan J M, Geddes N J, Flamant P D H and Flesia C 1998 *A global backscatter database for aerosols and cirrus clouds: Report to ESA 12510/97/NL/RE* p. 110, June

[68] Kent G S and Schaffner S K 1989 Comparison of 1μm satellite aerosol extinction with CO_2 lidar backscatter *SPIE* **1181** 252–259

[69] Rosen J M 1991 A comparison of measured and calculated optical properties of atmospheric aerosols at infrared wavelengths *J. Geophys. Res.* **96** 5229–5235

[70] Chudamani S, Spinhirne J D and Clarke A D 1996 Lidar aerosol backscatter cross sections in the 2μm near-infrared wavelength region *Appl. Opt.* **35** 4812–4819

[71] Phillips M W, Hannon S, Henderson S W, Gatt P and Huffaker R M 1997 Solid-state coherent lidar technology for space-based wind measurement *SPIE* **2965** 68–75

[72] Winker D M, Couch R H and McCormick M P 1996 An overview of LITE: NASA's lidar in space technology experiment *Proc. IEEE* **84** 164–180

[73] Kent G S, Poole L R and McCormick M P 1986 Characteristics of arctic polar stratospheric clouds as measured by airborne lidar *J. Atmos. Sci.* **43** 20

[74] Powell K A, Trepte C R and Kent G S 1996 *Observation of Saharan dust by LITE* Advances in Atmospheric Remote Sensing with Lidar(Berlin: Springer)

[75] Bowdle D A and Menzies R T 1998 *Aerosols backscatter at 2 μm: modelling and validation for SPARCLE and follow-on missions* In NOAA Working Group Meeting on Space-Based Lidar Winds, Key West, FL, January 20–22

[76] Srivastava V, Rothermel J, Clarke A D, Spinhirne J D, Menzies R T, Cutten D R, Jarzembski M, Bowdle D A and McCaul E W Jr 2001 Wavelength dependence of backscatter by use of aerosol microphysics and lidar data sets: application to 2.1-μm space-based and airborne lidars *Appl. Opt.* **40** 4759–4769

[77] Vaughan J M, Maryon R H and Geddes N J 2002 Comparison of atmospheric aerosol backscattering and air mass back trajectories *Meteorol. Atmos. Phys.* **79** 33–46

[78] Carrier L W, Cato G A and von Essen K J 1967 The backscattering and extinction of visible and infrared radiation by selected major cloud models *Appl. Opt.* **6** 1209–1216

[79] Evans B T N and Fournier G R 1990 Simple Approximations to extinction efficiency valid over all size parameters *Appl. Opt.* **29** 4666–4670

[80] Chylek P and Damiano P 1992 Polynomial approximation of the optical properties of water clouds in the 8–12μm spectral region *J. Appl. Meteorol.* **31** 1210–1218

[81] Banakh V A, Smalikho I N, Kopp F and Werner C 1995 Representativity of the wind measurements by a CW Doppler lidar in the atmospheric boundary layer *Appl. Opt.* **34** 2055–2067

[82] Harris M, Constant G and Ward C 2001 Continuous wave bistatic laser Doppler wind sensor *Appl. Opt.* **40** 1501–1506

[83] Karlsson C, Olsson F, Letalick D and Harris M 2000 All-Fibre multifunction continuous-wave 1.55μm coherent laser radar for range, speed, vibration and wind measurements *Appl. Opt.* **39** 3716–3726

[84] Jarzembski M A, Srivastava V and Chambers D M 1996 Lidar calibration technique using laboratory-generated aerosols *Appl. Opt.* **35** 2096–2108

[85] Vaughan J M and Forrester P A 1989 Laser Doppler velocimetry applied to the measurement of local and global wind *Wind Eng.* **13** 1–15

[86] Clemesha B R, Kirchhoff V W J H and Simonich D M 1981 Remote measurement of tropospheric winds by ground based lidar *Appl. Opt.* **20** 2907–2910

[87] Schols J L and Eloranta E W 1992 The calculation of area-averaged vertical profiles of the horizontal wind velocity from volume imaging lidar data *J. Geophys. Res.* **97** 18395–18407

[88] Pirronen A K and Eloranta E W 1995 Accuracy analysis of wind profiles calculated from volume imaging lidar data *J. Geophys. Res.* **100** 25559–25567

[89] Willetts D V and Harris M R 1982 An investigation into the origin of frequency sweeping in a hybrid TEA CO2 laser radar *J. Phys.* **15** 51–67

[90] Willets D V and Harris M R 1983 Scaling laws for the intrapulse frequency stability of an injection mode selected TEA CO_2 laser *IEE J. Quant. Electron.* **QE-19** 810–814

[91] Wilson D J, Constant G D J, Foord R and Vaughan J M 1991 Detector performance studies for CO_2 laser heterodyne systems *Infrared Phys.* 1 **31** 109–115

[92] Oh D, Drobinski P, Salamitou P and Flamant P H 1996 Optimal local oscillator power for CMT photo-voltaic detector in heterodyne mode *Infrared Phys. Technol.* **37** 325–333

[93] Lottman B T and Frehlich R G 1997 Evaluation of coherent Doppler lidar velocity estimators in nonstationary regimes *Appl. Opt.* **36** 7906–7918

[94] Rye B J and Hardesty R M 1989 Time verification and Kalman filtering techniques for Doppler lidar velocity estimation *Appl. Opt.* **28** 879–891

[95] Rye B J and Hardesty R M 1993 Discrete spectral peak estimation in incoherent backscatter heterodyne lidar. I: spectral accumulation and the Cramer-Rao lower bound II Correlogram accumulation *IEEE Transactions of Geoscience and Remote Sensing* **31** 1

[96] Rye B J and Hardesty R M 1997 Estimate optimisation parameters for incoherent backscatter heterodyne lidar *Appl. Opt.* **36** 36

[97] Dabas A M, Drobinski P, Flamant P H and Dabas A M 1999 Adaptive Levin filter for frequency estimate of heterodyne Doppler lidar returns: recursive implementation and quality control *J. Atmos. Oceanic Technol.* **16** 361–372

[98] Rye B J 2000 Estimate optimisation parameters for incoherent backscatter lidar including unknown return signal bandwidth *Appl. Opt.* **39** 6068–6096

[99] Eberhard W L, Cupp R E and Healy K R 1989 Doppler lidar measurement of profiles of turbulence and momentum flux *J. Atmos. Oceanic Technol.* **6** 809–819

[100] Post M J and Neff W D 1986 Doppler lidar measurement of winds in a narrow mountain valley *Bull. Am. Meteorol. Soc.* **67** 274

[101] Neiman R J, Hardesty R M, Shapiro M A and Cupp R E 1998 Doppler lidar observations of a downslope windstorm *Mon. Weather Rev.* **116** 2265–2275

[102] Olivier L D and Banta R M 1991 *Doppler lidar measurements of wind flow and aerosol concentration at the Grand Canyon, Technical Digest* Coherent Laser Radar: Technology and Applications July 8–12 1991 (Snowmass, Colorado: Optical Society of America)

[103] Hardesty R M and Huffaker R M 1996 Remote sensing of atmospheric wind velocities using solid state and CO2 coherent laser systems *Proc. IEEE* **84** 181–204

[104] Pearson G N and Collier C G 1999 A pulsed coherent CO2 lidar for boundary layer meteorology *QJR Meteorol. Soc.* **125** 2703–2721

[105] Hawley J G, Targ R, Henderson S W, Hale C P, Kavaya M J and Moerder D 1993 Coherent launch-site atmospheric wind sounder: theory and experiment *Appl. Opt.* **32** 4557–4568

[106] Frehlich R 1995 Comparison of 2- and 10-μm coherent Doppler lidar performance *J. Atmos. Oceanic Technol.* **12** 2

[107] Pearson G N, Roberts P J, Eacock J R and Harris M 2002 Analysis of the performance of a coherent pulsed fibre lidar for aerosol backscatter applications *Appl. Opt.* **41** 6442–6450

[108] McKay J A 1998 Modelling of direct detection Doppler wind lidar *Appl. Opt.* **37** 27

[109] Bruneau D and Pelon J 2003 Simultaneous measurements of particle backscattering and extinction coefficients and wind velocity by lidar with a Mach-Zehnder interferometer: principle of operation and performance assessment *Appl. Opt.* **42** 1101–1114

[110] Abreu V J, Barnes J E and Hays P B 1992 Observations of winds with an incoherent lidar detector *Appl. Opt.* **31** 4509–4514

[111] McGill M J, Skinner W R and Irgang T D 1997 Analysis techniques for the recovery of winds and backscatter coefficients from a multiple channel incoherent Doppler lidar *Appl. Opt.* **36** 1253–1268

[112] Rees D, Vyssogorets M, Meredith N P, Griffin E and Chaxel Y 1996 The Doppler wind and temperature system of ALOMAR lidar facility: overview and initial results *J. Atmos. Terr. Phys.* **58** 1827–1842

[113] Garnier A and Chanin M L 1992 Description of a Doppler Rayleigh LIDAR for Measuring Winds in the Middle Atmosphere *Appl. Phys.* **B55** 35–40

[114] Gentry B and Korb C L 1994 Edge technique for high accuracy Doppler velocimetry *Appl. Opt.* **33** 5770–5777

[115] Flesia C and Korb C L 1999 Theory of the double edge molecular technique for Doppler lidar wind measurement *Appl. Opt.* 3 **38** 432

[116] Souprayen C, Garnier A, Hertzog A, Hauchecorne A and Porteneuve J 1999 Rayleigh Mie Doppler wind lidar for atmospheric measurements. I. Instrumental setup, validation and first climatological results: II Mie scattering effect, theory and calibration *Appl. Opt.* **38** 2410–2421

[117] Delaval A, Flamant P H, Aupierre M, Delville P, Loth C, Garnier A, Souprayen C, Bruneau D, Le Rille D, Wilson R, Vialle C, Rees D, Vaughan J M and Hardesty R M 2001 Intercomparison of wind profiling instruments during the VALID field campaign *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon Editions de l'Ecole Polytechnique pp 101–104

[118] Pal S R, Ryan J S and Carswell A I 1978 Cloud reflectance with laser beam illumination *Appl. Opt.* **17** 2257

[119] Kent G S, Poole L R and McCormick M P 1986 Characteristics of arctic polar stratospheric clouds as measured by airborne lidar *J. Atmos. Sci.* **43** 20

[120] Hall F F, Cupp R E and Troxel S W 1988 Cirrus cloud transmittance and backscatter in the infrared measured with a CO2 lidar *Appl. Opt.* **27** 2510–2539

[121] Krichbaumer W, Mehneet A, Halldorsson T H, Hermann H, Haering R, Streicher J and Werner C H 1993 A diode-pumped Nd: YAG lidar for airborne cloud measurements *Opt. Laser Technol.* **25** 283–287

[122] Keeler R J, Serafin R J, Schwiesow R L, Lenschow D H, Vaughan J M and Woodfield A A 1987 An airborne laser air motion sensing system. Part I: concept and preliminary experiment *J. Atmos. Oceanic Technol.* 1 **4** 113–137

[123] Woodfield A and Vaughan J M 1983 Airspeed and windshear measurement with an airborne CO2 laser *Int. J. Aviation Safety* **1** 207–224

[124] Woodfield A and Vaughan J M 1983 Airspeed and windshear measurement with an airborne CO2 CW laser *AGARDo-graph* **272** 7.1–7.17

[125] Woodfield A and Vaughan J M 1984 Using an airborne CO2 laser for free stream airspeed and windshear measurements *AGARD Conf. Proc.* **373** 22.1–22.18

[126] Targ R, Kavaya M, Milton Huffaker R and Bowles R L 1991 Coherent lidar airborne windshear sensor: performance evaluation *Appl. Opt.* **30** 2013–2026

[127] Morbieu B, Combe H, Mandle J 1993, ALEV3, a 3-axis CO2 CW anemometer for aircraft certification, *Proc. 7th Conf. On Coh. Las. Rad.*, Appl. Tech., Paris, Paper MB1.

[128] Smart A E 1991 Velocity sensor for an airborne optical air data system *AIAA: J. Aircraft* 3 **28** 163–164

[129] Smart A E 1992 Optical velocity sensor for air data applications *Opt. Eng.* 1 **31** 166–173

[130] Werner C, Flamant P H, Reitebuch O, Kopp F, Streicher J, Rahm S, Nagel E, Klier M, Hermann H, Loth C, Delville P, Drobinski P, Romand B, Boitel C, Oh D, Lopez M, Meissonnier M, Bruneau D and Dabas A M 2001 WIND infrared Doppler lidar instrument *Opt. Eng.* **40** 115–125

[131] Reitebuch O, Werner C, Leike I, Delville P, Flamant P H, Cress A and Englebart D 2001 Experimental validation of wind profiling performed by the airborne 10μm-Heterodyne Doppler lidar WIND *J. Atmos. Oceanic Technol.* **18** 1331–1344

[132] Hallock J N 1991, Aircraft wake vortices: an annotated bibliography (1923-1990), Report no. DOT-FAA-RD-90-30, DOT-VNTSC-FAA-90-7 (available through: National Technical Information Service, Springfield, Virginia 22161, USA).

[133] Koepp F 1994 Doppler lidar investigations of wake vortex transport between closely spaced parallel runways *AIAA J.* **32** 805–812

[134] Constant G D J, Foord R, Forrester P A and Vaughan J M 1994 Coherent laser radar and the problem of wake vortices *J. Mod. Opt.* **41** 2153–2174

[135] Vaughan J M, Steinwall K O, Werner C and Flamant P H 1996 Coherent laser radar in Europe *Proc. IEEE* **84** 205–226

[136] Harris M, Young R I, Kopp F, Dolfi A and Cariou J-P 2002 Wake vortex detection and monitoring *Aerosp. Sci. Technol.* **6** 325–331

[137] Hannon S M and Thomson J A 1994 Aircraft wake vortex detection and measurement with pulsed solid-state coherent laser radar *J. Mod. Opt.* **41** 2175–2196

[138] Koepp F 1999 Wake vortex characteristics of military-type aircraft measured at Airport Oberpfaffenhofen using the DLR Laser Doppler Anenometer *Aerosp. Sci. Technol.* **4** 191–199

[139] Joseph R, Dasey T, Heinrichs R 1999, Vortex and meteorological measurements at Dallas/Fort Worth Airport, AIAA 99-0760.

[140] Greenwood J S and Vaughan J M 1998 Measurements of aircraft wake vortices at Heathrow by laser Doppler velocimetry *Air Traffic Control Quarterly* **6** 179–203

[141] Vaughan J M 1998, Wake vortex investigations at Heathrow airport, London, Nouvelle Revue d'Aeronautique et d'Astronautique, 2, 116–121.

[142] Harris M, Huenecke K and Huenecke C 2000 Aircraft wake vortices: a comparison of wind-tunnel data with field-trial measurements by laser radar *Aerosp. Sci. Technol.* **4** 363–370

[143] Vaughan J M and Harris M 2001 Lidar measurement of B747 wakes: observation of a vortex within a vortex *Aerosp. Sci. Technol.* **5** 409–411

[144] Turner G P, Padfield G D and Harris M 2002 Encounters with aircraft vortex wakes: the impact on helicopter handling qualities *J. Aircraft* **39** 839–849

[145] Grant W B 1991 Differential absorption and Raman lidar for water vapour profile measurements: a review *Opt. Eng.* 1 **30** 40–48

[146] Klett J D 1981 Stable analytical inversion solution for processing lidar returns *Appl. Opt.* 1 **20** 211–215

[147] Ismail S and Browell E V 1989 Airborne and spaceborne lidar measurements of water vapour profiles: a sensitivity analysis *Appl. Opt.* 17 **28** 3603–3615

[148] McDermid I S, Godin S M and Lindqvist L O 1990a Ground-based laser DIAL system for long-term measurements of stratospheric ozone *Appl. Opt.* 25 **29** 3603–3612

[149] McDermid I S, Godin S M and Walsh T D 1990b Lidar measurements of stratospheric ozone and intercomparisons and validation *Appl. Opt.* 33 **29** 4914–4923

[150] Hardesty R M 1984 Coherent DIAL measurement of range-resolved water vapour concentration *Appl. Opt.* 15 **23** 2545–2553

[151] Grant W B, Margolis J S, Brothers A M and Tratt D M 1987 CO_2 DIAL measurements of water vapour *Appl. Opt.* 15 **26** 3033–3042

[152] Grant W B 1989 Mobile atmospheric in pollutant mapping (MAPM) system: a coherent CO2 DIAL system, Laser Applications in Meteorology and Earth and Atmospheric Remote Sensing *Proc. SPIE* **1062** 172–190

[153] Ridley K D, Pearson G N and Harris M 2001 Influence of speckle correlation on coherent DIAL with an in-fibre wavelength multiplexed transceiver *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon Editions de l'Ecole Polytechnique pp 93–96

[154] Godin S, Marchand M and Hauchecorne A 2001 Study of the influence of the Arctic polar vortex erosion on mid-latitude from ozone lidar measurements at OHP (44°N, 6°E) *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon (Editions de l'Ecole Polytechnique) pp 385–387

[155] Butler C F, Browell E V, Grant W B, Brackett V G, Toon O B, Burris J, McGee T, Schoeberl M and Mahoney M J 2001 Polar stratospheric cloud characteristics observed with airborne lidar during the SOLVE campaign *Advances in Laser Remote Sensing* eds A Dabos, C Loth and J Pelon (Editions de l'Ecole Polytechnique) pp 397–400

[156] Gerard E and Pailleux J 2001 Role of water vapour in Numerical Weather Prediction models *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon (Editions de l'Ecole Polytechnique) pp 285–288

[157] Riedinger E, Keckhut P, Hauchecorne A, Collado E and Sherlock V 2001 Monitoring water vapour in the mid-upper troposphere using ground-based Raman lidar *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon (Editions de l'Ecole Polytechnique) pp 313–315

[158] Sherlock V, Hauchecorne A and Lenoble J 1999 Methodology for the independent calibration of Raman backscatter water-vapour lidar systems *Appl. Opt.* **38** 5817–5837

[159] Ferrare R A, Whiteman D N, Melfi S H, Evans K D, Schmidlin F J and O'C Starr D 1995 A Comparison of Water Vapour Measurements made by Raman Lidar and Radiosondes , *J. Atmos. Oceanic Technol.* **6** 1177–1195

[160] Evans K D, Demoz B B, Cadirola M P and Melfi S H 2001 A new calibration technique for Raman Lidar Water Vapour Measurements *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon Editions de l'Ecole Polytechnique pp 289–292

[161] Whiteman D and Melfi S H 1999 Cloud liquid water, mean droplet radius and number density measurements using a Raman lidar *J. Geophys. Res.No* D24 **104** 31411–31419

[162] Ansmann A, Riebesell M and Weitkamp C 1990 Measurement of atmospheric aerosol extinction profiles with a Raman lidar *Opt. Lett.* **15** 746–748

[163] Ansmann A, Wandinger U, Riebesell M, Weitkamp C and Michaelis W 1992 Independent measurement of extinction and backscatter profiles in cirrus clouds by using a combined Raman elastic-backscatter lidar *Appl. Opt.* **31** 7113–7131

[164] Schumacher R, Neuber R, Herber A and Rairoux P 2001 Extinction profiles measured with a Raman lidar in the Arctic troposphere *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon Editions de l'Ecole Polytechnique pp 229–232

[165] Uthe E E 1991 Elastic scattering, fluorescent scattering and differential absorption airborne lidar observations of atmospheric tracers *Opt. Eng.* 1 **30** 66–71

[166] Eberhard W L and Chen Z 1989 Lidar discrimination of multiple fluorescent tracers of atmospheric motions *Appl. Opt.* **28** 2966–3007

[167] She C Y, Latifi H, Yu R J, Alvarez R J, Bills R E and Gardner C S 1990 Two frequency lidar technique for mesospheric Na temperature measurements *Geophys. Res. Lett.* **17** 929–932

[168] Gelbwachs J A 1994 Iron Boltzmann factor lidar: proposed new remote-sensing technique for mesospheric temperature *Appl. Opt.* **33** 7151–7156

[169] Von Zahn U, Gerding M, Hoffner J, McNeill W J and Murad E 1999 Fe Ca and K atom densities in the trails of Leonid and other meteors: strong evidence for differential ablation *Meteor. Planet Sci.* **34** 1017–1027

[170] Bills R E, Gardner C S and She C Y 1991 Narrowband lidar technique for sodium temperature and Doppler wind observations of the upper atmosphere *Opt. Eng.* 1 **30** 13–20

[171] Chu X, Gardner C S, Pan W and Papen G C 2001 Recent results from the University of Illinois Iron Boltzmann temperature lidar *Advances in Laser Remote Sensing* eds A Dabas, C Loth and J Pelon Editions de l'Ecole Polytechnique pp 413–416

[172] Melngailis I, Keicher W E, Freed C, Marcus S, Edwards B E, Sanchez A, Fan T Y and Spears D L 1996 Laser component Technology *Proc. IEEE* **84** 2

[173] Bufton J L, Garvin J B, Cavanaugh J F, Ramos-Izquierdo L, Clem T D and Krabill W B 1991 Airborne lidar for profiling of surface topography *Opt. Eng.* **30** 72–78

[174] Gardner C S 1982 Target signatures for laser altimeters; an analysis *Appl. Opt.* **4** 448–453

[175] Bufton J L 1989 Laser altimetry measurements from aircraft and spacecraft *Proc. IEEE* 3 **77** 463–477

[176] Garvin J B, Bufton J L, Krabill W B, Clem T D Airborne laser altimeter observations of Mount St. Helens volcano, October 1989, submitted to J. Volcanology and Geothermal Res., June 1990.

[177] Garvin J B, Bufton J L, Krabill W B, Clem T D, Schnetzler C C Airborne laser altimetry of craterform structures, submitted to IEEE Trans. Geo. Rem. Sens., June 1990.

[178] Steinvall O, Koppari K and Karlsson U 1993 Experimental evaluation of an airborne depth sounding lidar *Opt. Eng.* **32** 1307–1321

[179] Hoge F E, Wright C W, Krabill W B, Buntzen R R, Gilbert G D, Swift R N, Yungle J K and Berry R E 1988 Airborne lidar detection of subsurface oceanic scattering layers *Appl. Opt.* **27** 3969–3977

[180] Penny M F, Billard B and Abbot R H 1989 LADS — The Australian Laser airborne depth sounder *Int. J. Remote Sensing* **10** 1463–1497

[181] Lutormirski R F 1978 An analytical model for optical beam propagation through the maritime boundary layer *SPIE Ocean OpticsV* **160** 110–122

[182] Wood R and Appleby G 2003 *Earth and environmental sciences: satellite laser ranging* Article D7.1 Handbook of Laser Technology and Application(Bristol: IOPP)

[183] Plotkin H 1964 S66 laser satellite tracking experiment *Proceedings of the Quantum Electronics III Conference* (Newyork: Columbia University Press) pp 1319–1332

[184] Nuebert R, Grunwaldt L and Fisher H 2001 *The new SLR station of GFZ Potsdam: A status report* Proceedings of 12th International Workshop on Laser Ranging, Matera, Italy

[185] Schwartz J A 1990 Pulse spreading and range correction analysis for satellite laser ranging *Appl. Opt.* **29** 3597–3602

[186] Couch R, Rowland C, Ellis K, Blythe M, Regan C, Koch M R, Antill C W, Kitchen W L, Cox J W, De Lorme J F, Crockett S K, Remus R W, Casas J C and Hunt W H 1991 Lidar In-space Technology Experiment (LITE): NASA's first in-space lidar system for atmospheric research *Opt. Eng.* **30** 88–95

[187] McCormick M P, Winker D M, Browell E V, Coakley J A, Gardner C S, Hoff R M, Kent G S, Melfi S H, Menzies R T, Platt C M R, Randall D A and Reagan J A 1993 Scientific investigations planned for the Lidar In-space Technology Experiment (LITE) *Bull. Am. Meteorol. Soc.* **74** 205–214

[188] Megie G 1997 Differential Absorption Lidar *Laser Beams in Space: Application and Technology* ed J Bufton (New York: Marshal Dekker)

[189] Cohen S C, Degnan J J, Bufton J L, Garvin J B and Abshire J B 1987 The geoscience of laser altimetry/ranging systems *IEEE Trans. GeoSci. Rem. Sens. GE* 5 **25** 581–592

[190] Hardesty R M and Weber B F 1987 Lidar measurement of turbulence encountered by horizontal-axis wind turbines *J. Atmos. Oceanic Technol.* **4** 191–203

[191] Steinvall O, et al. 1981 Physics and Technology of Coherent Infrared Radar *SPIE* **300** 104

[192] Osche G R and Young D S 1996 Imaging laser-radar in the near and far infrared *Proc. IEEE* 2 **84** 103–125

[193] Stephan B and Metivier P 1987, Flight evaluation trials of a heterodyne CO2 laser radar, *Active Infrared Syst. Techn.*, SPIE-806, 110–118.

[194] Hogg G M, Harrison K and Minisclou S 1995 Nato Agard *Conf. Proc.* **563** 20

[195] ATLID Consultancy Group 1990. Backscatter Lidar, the potential of a space-borne lidar for operational meteorology, climatology and environmental research, ESA specialist publication, SP — 1121.

[196] Atmospheric Dynamics Mission 1999, *The Four Candidates Earth Explorer Core Missions.* ESA Reports for Mission Selection, ESA-SP 1233 (4). Noordwijk ESA.

Further reading

Valuable accounts of topics in light scattering and laser radar may be found in the following conference proceedings, reviews, compendia etc.

Killinger D K and A Moorodian eds 1983 *Optical and Laser Remote Sensing* (New York: Springer)

Measures R M ed 1988 *Laser Remote Chemical Analysis* (New York: Wiley)

Vaughan J M 1989 *The Fabry-Perot Interferometer — History, Theory, Practice and Applications* Adam Hilger Series (Bristol: IOPP)

Jelalian A V 1992 *Laser Radar Systems* (Boston: Artech)

Journal of Modern Optics 41 (11), November 1994, 2063–2196 (containing a Special Section of 10 papers on Coherent Laser Radar, with short preface by J M Vaughan).

Proceedings of the IEEE 84 (2), February 1996, 99–320 (Special Issue of 8 papers on Laser Radar, ed. A V Jelalian).

Ansmann A, R Neuber, P Rairoux and U Wandinger eds 1996 Advances in atmospheric remote sensing with Lidar *International Laser Radar Conference (ILRC), Berlin, Germany, July 22–26* (Berlin: Springer)

Dabas A, C Loth and J Pelon eds 2001 Advances in Laser Remote Sensing *Selected Papers presented at the 20th International Laser Radar Conference (ILRC), Vichy, France 2001* Editions de l'Ecole Polytechnique

Conference Proceedings of the Biennial International Laser Radar Conferences (ILRC).

Technical Digests of the Biennial Coherent Laser Radar Meetings (CLRM).

C3.3
Military optoelectronics

Hilary G Sillitto

C3.3.1 Introduction

C3.3.1.1 Purpose

This chapter surveys the military needs for and applications of optoelectronics, and illustrates these with examples. The intention is to give the interested reader a conceptual overview of this very wide subject and to provide references for analytical detail.

C3.3.1.2 The military value of electro-optics

Three key benefits of electro-optical technology have led to its widespread adoption by the military. They are:

- high angular resolution through a small aperture, because of the short operating wavelength—making useful electro-optic (EO) systems easy to package on a wide range of platforms;
- twenty-four hour operation—passive night vision and thermal imaging systems can 'turn night into day';
- the familiarity of the 'display metaphor'—images from EO systems look like the images we see with our eyes, which makes them easy to interpret and makes it easy to train operators.

C3.3.1.3 Themes and structure of this chapter

This chapter starts with a brief historical perspective.

The key building blocks of modern military EO systems are imaging and laser subsystems. These and other key enabling technologies are then described.

The uniquely demanding characteristics of the military operating environment are then outlined, since these strongly influence the engineering of optoelectronic technology into military systems.

The roles, functions, operational characteristics, technology and representative examples of the main classes of military optoelectronic systems are then reviewed.

The chapter concludes with an assessment of the operational impact of optoelectronic technologies, and tempts fate by outlining the characteristics of current trends in the field which give us insight into how it may evolve in the future.

The author is attempting to provide a neutral overview of the subject, and unreservedly apologizes for the 'English-speaking/North-west European' bias that probably remains.

C3.3.1.4 The EO environment

Military EO systems are remote sensing systems. Their performance is critically influenced by 'the EO environment': atmospheric transmission and scatter, target and clutter signatures, background emission and reflection, smoke and cloud, fires, and atmospheric scintillation.

Figure C3.3.1 shows, superimposed on a spectrum from 200 nm to 13 μm, six primary influences on military visual and IR system design. They are:

- *The transmission of the atmosphere*: the main atmospheric 'windows' are 0.4–1.6, 1.8–2.5, 3–5 μm (split by the CO_2 absorption band at 4.2), and 8–12 μm; there is much fine structure due to molecular absorption lines within these 'windows', and a water vapour absorption continuum in the 8–13 μm band;
- *The solar radiation spectrum*, approximating to a 6000 K black body and peaking at about 550 nm when plotted on a wavelength scale (but at 880 nm when plotted on a frequency scale—see reference [28] for a discussion of the common assumption that the human eye response is matched to the peak of the solar spectrum);
- *The human eye response* from 400 to 700 nm;
- *The silicon detector response* from 400 to 1100 nm;
- *The hot CO_2 emission band* surrounding the absorption band at 4.2 μm, which is the main source of the strong 3–5 μm signature of exhaust plumes and fires;
- *The 300 K black body curve* representing the emission of natural objects at ambient temperature, showing the dominance of emission in the 8–12 and 4–5 μm regions.

These six factors dominate the design and selection of military EO systems.
Other significant atmospheric factors include:

- Atmospheric scatter: rayleigh scattering has an inverse fourth power relationship with wavelength, explaining why longer wavelength systems are less susceptible to haze and smoke;

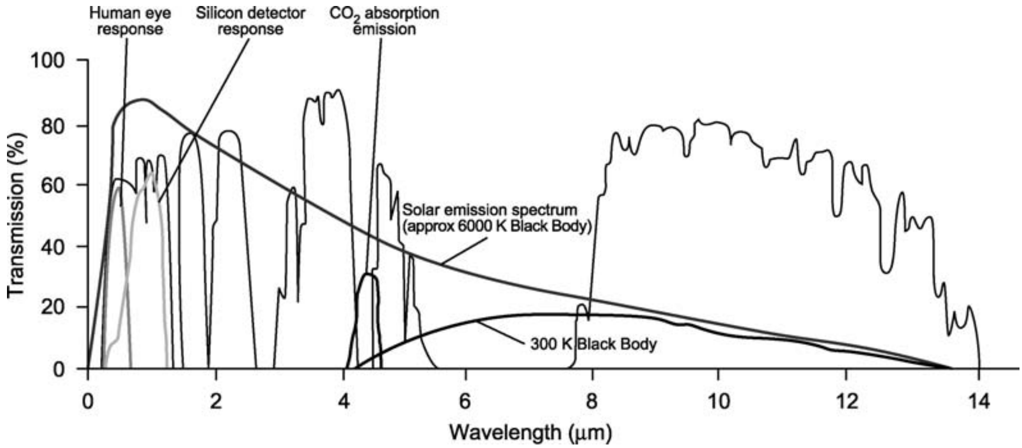


Figure C3.3.1. The primary spectral influences on military EO system design.

- Atmospheric scintillation, which limits the resolution of high magnification imaging systems and the divergence and beam uniformity of lasers working through atmospheric paths; scintillation is caused by density and humidity fluctuations in the air, and is much stronger near the ground;
- Clouds, which except when very thin are essentially opaque to EO systems;
- Screening smoke, which is discussed later;
- Atmospheric ducting due to refractive index gradients near the ground or sea surface, which can slightly increase the horizon range of low horizontal paths by causing the light to bend to follow the earth's curvature.

Target and clutter phenomena are dominated by reflectance and emission of the surface. These vary with wavelength, sometimes quite markedly; for example natural vegetation contains chlorophyll which reflects strongly in the near IR from just above 700 nm. Shiny metallic surfaces have low IR emissivity and therefore reduce the apparent temperature of for example an aircraft; but of course reflect sunlight strongly, making the aircraft easier to detect in the visual band. Hot fires and exhaust gases are extremely prominent in the 3–5 because of the rotational CO_2 emission bands around 4.2 μm, making 3–5 μm imagers less useful for ground-to-ground applications in intense battle conditions than their performance in good conditions would suggest. Measures taken to reduce signature in one band will often increase signatures in another.

C3.3.2 Historical perspective

C3.3.2.1 Myths, legends and fantasies—'from the Greeks to Star Wars'

Man has always fantasized about optical system concepts which would give him an advantage in warfare.

'And I' said Athena (in the legend of Perseus and the Gorgon) in her calm, sweet voice, 'will lend you my shield with which I dazzle the eyes of erring mortals who do battle against my wisdom. Any mortal who looks upon the face of Medusa is turned to stone by the terror of it; but if you look only on her reflection in the shield, all will be well' [1].

This use of bright light to dazzle the enemy, the idea of a 'death ray' weapon, and the use of indirect viewing optics to mitigate its effect, all seem to anticipate by several thousand years a number of the 20th century's threats, opportunities and solutions.

In H.G. Wells' science fiction novel 'The War of the Worlds' [2], the Martians apparently used some sort of directed energy weapon, which we would now assume was a high energy laser. Possibly inspired by this, Britain's Tizard commission in the 1930s [3] set a goal of a death ray, which was not achieved, but led to the successful UK radar programme.

In the 1980s Sci-fi film, 'Robocop' used a helmet-mounted head-up information display, and was alleged in a TV documentary to have contributed inspiration to at least one of the Future Soldier Technology programmes in the late 1990s.

In the Star Trek television series, it takes little imagination to equate phasors with laser weapons; while the 'Star Wars' series of films gave its name to Reagan's Strategic Defence Initiative, and anticipated holographic 'video-conferencing'.

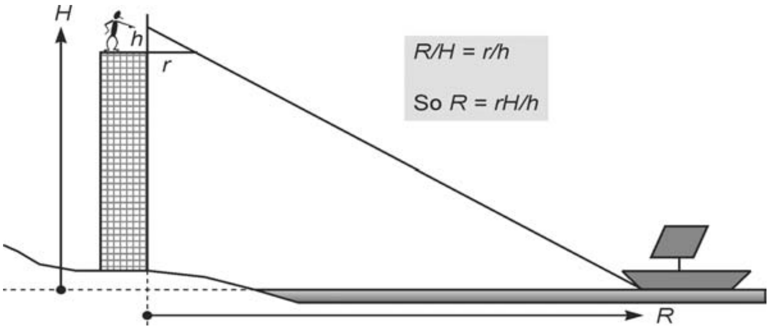


Figure C3.3.2. Thales of Miletus, 500 BC: how to measure the range to a ship from a tower.

C3.3.2.2 Optical systems and methods—'from the Greeks to World War 2'

Optical instruments and knowledge were used since Greek times in warfare. Around 500 years BC, Thales of Miletus proposed geometry allowing the range of ships to be estimated from a tall tower (figure C3.3.2).

The Heliograph has been used since ancient times for semi-covert line of sight signalling.

During the age of exploration from 1400 to 1900 AD, all civilizations placed great importance on navigation by sun and stars. This drove the development of optical instruments and precision navigation techniques—used for surveillance, long range identification, and (in conjunction with signal flags) for signalling.

Around 1900, the Boer War showed the importance and value of camouflage (signature suppression). Boer commandos used camouflage, mobility and modern rifles to negate British tactics used since the Napoleonic era. The British army introduced khaki uniforms for colonial campaigns in 1880, but for home service only in 1902 [27]. In the days when official war artists rather than CNN formed the public images of overseas warfare, artistic licence had perhaps veiled the switch to khaki from the British public; did the army used the Boer War experience as an 'excuse' to introduce the change at home?

In the late 19th century, Zeiss in Germany and Barr and Stroud in the UK developed optical rangefinders. By World War 1, Zeiss was recognized as the pre-eminent optical instrument manufacturer in Europe. Trench periscopes and binoculars were required in vast numbers during the war. The UK could no longer obtain Zeiss instruments, and faced a critical skills shortage. This led to the formation of the Imperial College Optics Group to develop an indigenous capability. Similarly, in the US, Frankford Arsenal played a key part in the development of military optical instruments [21]. All of these organizations and companies have featured strongly in the recent optoelectronics revolution.

C3.3.2.3 The technology revolution—lasers, IR and signal processing, 1955–1980

There has always been a close link between scientific and technological progress and military demands. World War 2 involved the scientific community in the war in an unprecedented manner.

In the UK from 1936, the Tizard committee spawned many far-reaching technological innovations. These included: the development and exploitation of radar; IR aircraft detection [3]; a photoelectric fuze [7]; operational analysis; the Enigma decoding organization at Bletchley Park leading to the development of the first electronic computer; and the systematic use of scientific intelligence [3]. The US's crowning technical achievements included the Manhattan project to develop the nuclear bomb; the efficient industrialization and mass production of all sorts of military technology; and of particular significance for optoelectronics, the development of the communications theory which underpins all modern signal processing techniques (Shannon, Bell labs). Germany made tremendous strides in many areas, including rocket and jet propulsion, and principles of modular design allowing cheap and distributed mass production of aircraft and tanks. During World War 2 the US and Germany developed and tested the first EO guided missiles, probably using PbS. Britain, Germany and the US all developed active near-IR night sights [21].

The subsequent development of military optoelectronics benefited from many aspects of this progress, and notably from the proof that technology managed at a strategic level could win wars.

Night vision was an important technological innovation in the immediate post war years. But three key breakthroughs after the war transformed optoelectronics from a supporting to a key military technology: in signal processing, IR, and lasers. The transistor was invented in 1948, the first forward-looking infrared (FLIR) was demonstrated in 1956 [11], and the first laser in 1960. These led to a whole new class of integrated EO systems, which developed rapidly in the 1970s and 1980s as the Cold War protagonists sought to achieve and maintain superiority. Government labs played a key part in pushing the technology and integrating the first systems. In the UK, the Royal Radar Establishment was established in Malvern during the war and later expanded its remit to include EO, becoming successively the Royal Signals and Radar Establishment, the Defence Research Agency (Electronics Division), and now Qinetiq. In the US, the Night Vision Lab at Fort Belvoir was founded in 1962 to co-ordinate night vision development for the US military [21].

C3.3.2.4 Military electro-optics in use

In Vietnam, IR linescan was deployed on American Grumman Mohawk aircraft in the 1960s. They were credited with being ‘able to detect the heat where a lorry had been parked hours after it had driven away’. Laser designation was used operationally by 1969—a remarkably short deployment time from the laser’s invention in 1960—allowing single planes to hit difficult targets such as bridges that had survived months of intense conventional bombing. Night Vision scopes helped US soldiers defend themselves against Viet-Cong night attacks.

In the British/Argentinean Falklands (Malvinas) conflict in 1982, the American AIM-9L Sidewinder missiles, with cooled InSb detector and head-on attack capability, allowed Royal Navy Harrier fighters to shoot down nearly 30 Argentinian aircraft without loss. Over 80% of the AIM-9Ls launched hit their targets [19]. Night Vision equipment was deployed by both sides’ armies, but does not seem to have been used particularly effectively, nor to have significantly influenced operations. Laser designation allowed effective stand-off bombing attacks without exposing attack aircraft to defending fire.

In Afghanistan in the 1980s, Stinger shoulder launched surface to air missiles (SAMs) were used against Soviet attack helicopters, which in turn used TV and IR imaging systems and laser designators for ground attack. Anti-missile countermeasures were used to reduce vulnerability of aircraft to SAMs: Soviet aircraft dispensing flares while landing and taking off at Afghan airfields became a familiar sight on TV newsreel shots. Equivalent Soviet missiles (SA-16 etc) were developed. The Soviet Union maintained a very strong research base in EO technologies throughout the Cold War.

In the Gulf Conflict in 1991, thermal imaging allowed the coalition forces to operate day and night, with an overwhelming advantage at night. The unusually bad weather and thick cloud cover, exacerbated by smoke from burning oil fields, made passive night vision equipment used by both sides almost useless for much of the time, but had little or no effect on thermal imagers. This experience led to an increase in demand for basic low cost thermal imagers for military roles for which the cheaper passive night vision technology had previously been considered sufficient.

The conflict showed the public, perhaps for the first time, an awesome array of precision weapons, many of them EO guided. The public perception of the conflict was shaped by live TV broadcasts (EO technology again) from the war zone, beamed by satellite and cable TV into homes throughout the world.

The coalition campaign used concepts of ‘manoeuvre warfare’ developed late in the Cold War. These depend on a high degree of integration of command, control, communication and information (C3I). EO is a key technology for capturing and displaying this integrated information picture; and, in the form of fibre-optic communications links, for the secure transmission of high bandwidth data between fixed command posts. It seems that fibre-optic cables were a high priority target for Special Forces sabotage missions and precision air strikes during the Gulf War.

Of the many issues for technologists emerging from the conflict, four are important to this discussion:

- Combat identification, to reduce the risk of attacking one’s own side, is particularly difficult and important in coalition warfare, where equipment used by friendly armies may not be compatible, and friends and enemies may be using the same types of equipment; and doubly so in manoeuvre warfare, where there is no clear front line.
- Collateral damage—even with precision weapons, not all weapons hit the intended targets (typical figures for laser guided bombs are 70–80% success rate); and real-time public television from the war zone ensures that the attacked party can exploit such lapses for propaganda purposes.
- The effectiveness of attacks was often overestimated—this is a persistent trend at least since World War 2 and probably since warfare began;
- It took time, often too much time, to get aircraft and satellite reconnaissance imagery to the people on the ground who could make use of it.

C3.3.2.5 Military electro-optics today

So, 40 years after the invention of the laser, EO is a well-established military technology of proven value. The issue for the 21st century is not whether, but how best and how widely, to exploit optoelectronics in military applications.

Passive EO sensors, such as thermal imagers, allow operators to detect, recognize and ‘identify’ objects of interest. With the addition of a laser rangefinder and directional sensing, the object can also be ‘located’ relative to the operator. With the further addition of a navigation capability, it can be located in geographical co-ordinates. The information can be shared with others via communication links, which in some circumstances may be optoelectronic (free space or fibre); and displayed on situation displays which use optoelectronic technology. When integrated with a weapon system, EO systems provide fire control solutions and help to attack the target. With the advent of high energy laser ‘weapons’, which in early 2002 were being developed as missile defence systems, EO systems can be used also to attack the target directly.

So EO systems answer the basic questions: ‘where is it? what is it? and (within limits) who is it and what is it doing?’, contribute to a shared information picture to aid military decision making, and in some contexts close the loop by attacking targets directly. This can be summarized in a closed loop referred to as the ‘OODA loop’ or ‘Boyd cycle’ (figure C3.3.3).

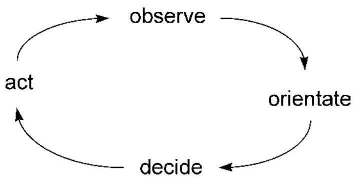


Figure C3.3.3. 'OODA loop' or 'Boyd cycle'.

C3.3.3 The key building blocks: imaging and laser technology

C3.3.3.1 Imaging

EO imaging systems convert optical radiation from the scene into electrical energy to allow it to be displayed in real time at a more convenient wavelength, brightness level, or location.

They are used for a wide spectrum of military and paramilitary tasks including surveillance, target acquisition, identification, weapon aiming, assessment, situation awareness, threat warning, pilot aid and covert driving. There is a correspondingly wide range of equipment and technology.

The wavelength range from 400 to about 1700 nm is referred to as 'visible and near infrared' (VNIR). VNIR image sensors depend on reflected light, using TV and image intensifier technology. Early active VNIR imaging systems used IR illuminators to augment natural light. Detection and recognition of objects of interest depends on the contrast between the object of interest and its surroundings, and on the resolution of the sensor. Modern high resolution TV technology is now allowing VNIR imaging systems to provide resolution approaching, but not quite matching, that obtained by the human eye through a high power magnifying sight of similar aperture and field of view (FOV).

Thermal imagers operate in the mid and far IR bands (3–5 and 8–12 μm, respectively) by sensing the black body radiation emitted by objects in the scene. Users can detect targets passively that are visually camouflaged or concealed in deep shadow, or when there is no ambient light. In particular, 8–12 μm imagers also see through smoke which blocks shorter wavelengths. In recent conflicts this has given armies and air forces equipped with thermal imaging an overwhelming operational advantage over those which were not.

In some applications the ability to see terrain features is important (flying and driving aids, general surveillance and orientation). In hot and humid conditions, thermal imagers may experience 'thermal wash-out', when everything in the scene is at the same temperature and there is very little thermal contrast. During and after rainfall, terrain contrast is very low in both bands. The 8–12 μm band suffers from water absorption in hot humid conditions; in these conditions, reflected sunlight may provide 'TV-like' pictures in the 3–5. In very cold conditions, by contrast, there is minimal black body emission in the 3–5, leading to wash-out in this band—tactical targets will still be detected, but details of surrounding terrain and of man-made structures may be suppressed.

Because of the longer wavelengths, identification range of thermal imagers is usually limited by the available aperture.

For these reasons, thermal imagers are often complemented by VNIR imaging or direct sighting systems, to allow improved identification range when weather and lighting provide adequate image contrast, and to provide a complementary sensing capability in thermal wash-out conditions.

The following paragraphs discuss in turn the technologies and integration issues involved in the four key imaging methods: image tubes; TV cameras; cooled photon-detecting thermal imagers; and uncooled thermal detecting thermal imagers.

Image tubes

Image tubes were first developed and saw limited use during World War 2. In the British EMI design, selenium photocathodes released electrons, which were accelerated by an intense electric field through a vacuum onto a photo-emissive zinc sulphide screen. Similar US devices used a phosphor screen [28]. Early devices had low sensitivity and required active illumination. They were referred to as ‘image convertors’ since they converted near IR radiation into visible light. Other designs used an electron lens. Since the electron lens inverts the image, simple objective and eyepiece lenses could be used with no need for image erecting prisms.

Gibson [21] describes subsequent developments in night vision well, mainly from a US perspective. Key breakthroughs were multi-stage amplification, which removed the need for active illumination; and the micro-channel plate, which allowed high amplification image tubes to be made short enough to fit into head mounted systems, and are much less susceptible to blooming effects from bright lights. ‘Generation 2+’ and ‘third generation’ image tubes are now widely used in night vision goggles (NVG) and light weapon sights. They offer high resolution imagery which is easy to interpret. They operate in the near IR (700–1000 nm), where grass and vegetation have high reflectivity, giving a good view of most types of natural terrain, and high contrast for most targets of military interest. Modern systems work in ‘overcast starlight’. In very overcast conditions and where terrain contrast is poor (for example snow or sand) the user’s spatial awareness can be augmented by near IR laser illumination [21]. This runs the risk of being detected by an enemy equipped with similar near-IR imaging systems.

Cockpit and other lighting can interfere with NVG operation. ‘NVG compatible’ lighting is filtered to be detected by NVGs without saturating them.

TV cameras

Early TV cameras used vidicon tubes. This technology was in military use in the late 1960s, with TV guided bombs being used in Vietnam [23]. Solid state image sensors (single-chip silicon photodetector arrays with CCD or MOS on-chip processing) were developed during the 1970s and 1980s and found wide commercial application, for example for security. These sensors were much smaller and more robust than vidicons, and made a TV camera capability much easier to package into military systems. They are becoming smaller and more sensitive; this reduces the optical aperture required to provide useful performance in a wide range of light levels. There is increasing concern about the laser threat to operators’ eyes, and an increased desire to mount EO systems in places which are difficult to provide direct optical channels for the crew. TV sensors are used much more widely in military systems than most observers expected 10–15 years ago.

The civil security and broadcast markets have led the demand for lower and lower light level capability in solid state sensors. Image intensifiers have been fitted to the front of TV sensors for many years to give low light TV capability. Image intensified CCD cameras provide similar performance to image intensifiers without the need to get the operator’s eye near the sensor. These are also the sensors usually used in laser-gated imaging systems. Emerging technologies such as electron beam CCD (EBCCD) offer the promise of ‘all light level’ TV performance in a package little bigger than a daylight TV camera.

Cooled photon-detecting thermal imagers

After the first successful demonstrations in the 1950s, thermal imaging started to shape the whole field of military EO.

TVs and most lasers use glass optics, and could be integrated into existing visual sighting systems. Glass does not transmit thermal IR radiation, so thermal imaging was incompatible with existing systems. As well as the detector technology itself, thermal imaging required a completely new family of optomechanical system architectures, and parallel developments in optical materials and coatings, lens design, test and calibration techniques, analogue and digital electronics, cryogenics and servo-mechanical systems.

Early developments, in the West at any rate, were almost exclusively in the 8–12 μm band. Cadmium mercury telluride (CMT) became the detector material of choice for first and second generation thermal imagers. CMT is a semi-conductor alloy whose band gap can be tuned by varying the material ratio. The band gap is matched to the photon energy of 8–12 μm radiation, of the order of 0.1 eV. When photons of this energy impinge on the detector electrons are excited from the valence to the conduction bands, and allow current to flow through the detector. This current is amplified and turned into a video signal which is fed to a display.

For this process to work and produce a usable image, the detector must not be swamped by electrons rising to the conduction band with their own thermal energy. CMT is normally cooled to about liquid nitrogen temperatures, 77 K. Early demonstration systems used liquid nitrogen to cool the detector, which was placed in a ‘Dewar flask’ with an IR window. Joule–Thomson cooling was widely used from the 1970s to the 1990s; nitrogen or air from a high pressure gas bottle or compressor is forced through a very small orifice at the end of a tube, placed just behind the detector, which is itself mounted within a vacuum enclosure. The expansion of the high pressure gas cools it to the required temperature. Because of the logistics, safety and maintenance issues associated with high pressure gas this method has now been largely superseded by Stirling cycle cooling engines which are coupled to the detector. The detectors are mounted within a vacuum dewar to reduce convection and avoid condensation.

We are trying to detect small differences in large photon fluxes. First and second generation systems scan the detector across the scene resulting in a modulation appearing on the detector output. This temporal modulation represents the spatial modulation in the scene. The large DC component is subtracted from the signal and the residual difference signal is then displayed as a TV-like image. Many first generation systems sent this modulated output to LEDs which were viewed through the same scanning system, ensuring that scanner errors did not affect the perceived picture geometry. Others produce a CCIR or NTSC video signal; in these systems scanner timing errors are less acceptable since they result in distortion of the picture.

One method for DC removal is to AC couple the detector output. This is cost-effective and worked well in the US Common Modules, of which many thousands were built. These systems used an oscillating mirror to scan a linear array of 60 or 180 detector elements across the scene. The output is AC coupled, amplified and either sent directly to an array of LEDs which is viewed off the back of the oscillating mirror, or processed through an electronic multiplexer to produce an electronic video signal.

Under some conditions of high scene contrast, AC coupling can generate artefacts from a small hot image which can wipe out a large part of the picture. In the UK, developments at RSRE to combat this problem led to a very high performance detector technology called the signal processing in the element (SPRITE), or Tom Elliot’s device (TED), named after the inventor. This is an array of eight parallel oblong detectors, each about 10 times as long as it was wide. The drift velocity of the electrons in the CMT is set by electronic biasing to match the image scan rate, providing an improved signal-to-noise ratio equivalent to that obtained by ‘time delay and integration’ (TDI) with discrete detectors. SPRITEs require a more complex ‘serial/parallel’ scanning method, typically with a high speed spinning polygon to generate the linescan and a small framing mirror oscillating at the video frame rate. The next stage of signal processing is simpler, with only eight amplification stages instead of 60 or 180, and performs amplification, DC restoration and channel equalization, resulting in a very stable and uniform picture. This technology is used in the British ‘Class 2 Thermal Imaging Common Module’ (TICM 2) system, and the Thales ‘IR-18’ family, and is operational in many thousand systems worldwide. The technology evolved to 16-element systems with digital processing giving outstanding resolution, sensitivity and image quality at high definition TV bandwidths.

Many ingenious optomechanical scanning methods were developed, striving to make the best use of the available detectors. However, the next major innovation was ‘second generation’ detectors which eliminated the need for complex 2-axis scanning and for many parallel sets of pre-amplifier electronics.

Second generation thermal imagers use a photovoltaic operating mode in which the incident photons cause a charge build up which changes the voltage on a capacitor. A focal plane multiplexer carries out time delay and integration for a linear array, and reads the signal off the focal plane serially. Common array sizes are 240, 288 or 480×4, or $768 \times n$ in the latest UK high performance system, STAIRS C. The signal from each pixel is amplified, digitized, and adjusted to compensate for response nonuniformities between detector elements. Modern systems provide a 12 or 14-bit digital output for processing, and 8-bit digital or analogue video. Automatic or manual gain/offset correction optimizes the dynamic range of the video output. These systems started to become available around 1990, the French Sofradir 288×4 detector being the first to be widely available in series production.

Second generation technology has split into two evolutionary paths. One is ‘cheaper smaller lighter’, as exemplified by the Thales ‘Sophie’ product, which put high performance thermal imaging into a binocular sized package suitable for use by individual soldiers for the first time. The other is ‘higher performance in the same package’, as exemplified by the US horizontal technology integration (HTI) programme and the UK STAIRS C system.

By 2000, the next step in simplification of high performance thermal imaging systems is starting to appear in the market. Two-dimensional arrays working in the 8–12 μm band are being developed and entering low rate production based on two technologies. The first is CMT, an incremental evolution from second generation devices. Efforts are being made to allow operation at higher temperatures to reduce the demands on cooling engine performance. The second, completely new, technology is quantum wells. These use molecular beam epitaxy to build a precisely tailored atomic structure which controls the electron levels by quantum confinement. The spectral band is much narrower than for CMT; this narrower spectral response can be exploited to give a number of system design benefits. Operating temperatures are similar to or colder than the 77 K of first and second generation systems. At the time of writing the jury is still out as to which if either route will dominate the market.

Most cooled 3–5 μm systems use one of the three detector technologies. Indium antimonide, a compound rather than an alloy, gives very good intrinsic uniformity and response out to about 5.5 μm. CMT with a different composition optimized for 3–5 allows the long-wave cut-off to be brought in to a shorter wavelength. Platinum silicide exploits silicon TV methods to provide very uniform imagery but with lower quantum efficiency (about 1% compared with about 60% for InSb and CMT). Staring systems of TV resolution are now widely available. There is resistance to 3–5 μm systems for ground to ground applications in most NATO armies because of the better performance of 8–12 μm systems in extreme battlefield conditions of fires and smoke. However, the lower cost and better resolution through a given aperture is leading to increasing adoption of 3–5 systems where the smoke and fire argument does not dominate the choice.

Uncooled ‘thermal detection’ thermal imagers

Uncooled thermal imagers use a variety of physical principles to generate electric current or charge as a result of temperature difference. Since they respond to temperature difference, they need not be cooled; this reduced the cost and complexity compared with cooled systems. They do need fast optics and large detector elements to achieve good sensitivity, so lens sizes for narrow fields of view become large. This limits their applicability to shorter range systems.

Most uncooled technologies respond to high frequency changes in temperature, and use a chopper to modulate the image of the scene. If the chopper is removed they can detect moving targets in a static scene by using the motion for modulation—the principle used in passive IR burglar alarms.

The first uncooled imaging technology was the pyroelectric vidicon—essentially a vidicon tube with the phosphor replaced by a thin pyroelectric membrane. The electron beam was used to 'read' the charge built up on the back surface of the membrane. These membranes were not mechanically robust, and would split if irradiated by a focused CO_2 laser!

The first 'solid state' thermal arrays used pyroelectric material placed in contact with a readout circuit similar in principle to that used by a solid state TV camera. Early devices suffered from low resolution, nonuniformity and microphony. These issues have been progressively resolved through a number of generations of process and material technology [22].

Uncooled thermal imagers are now in service in several western armies as weapon and observation sights. Spin-out technology is in the market for automotive night vision applications.

C3.3.3.2 Characterizing imaging system performance

Visual 'identification' criteria

Target classification with EO systems traditionally depends on the operator's training. The 'Johnson criteria' are based on experiments carried out with military operators looking at real or synthetic targets using real or simulated imaging systems. The criteria are:

Task	Cycles resolved on the target for 50% confidence level of DRI in a single observation
Detection	1–1.5
Recognition	3–4
Identification	6–7

Modelled or measured sensor characteristics are used to calculate the minimum resolved temperature difference (MRTD), which is plotted against spatial frequency. These criteria are then used along with an assumed target size and temperature difference or contrast to predict 'DRI ranges' for a particular sensor against a particular target.

This procedure is effective at providing a common benchmark to compare the relative performance of different system designs. It is less effective at predicting actual operational performance: wise suppliers quote minimum figures in their performance specifications, so most systems are better than their specification; and wise customers specify a conservative signature for the calculation, so most targets have higher contrast or temperature than is used in the model. The net effect is that most thermal imager systems are operationally useful at much greater ranges than calculated DRI figures suggest.

However, it is also important to understand that 'identification' specifically refers to the ability of the observer to choose between a limited class of known target types, and refers to 'identification of vehicle (or aircraft) type'—for example F-16 rather than MiG-29 aircraft, T-72 rather than M-1 tank. It bears no relation to the different and vital task of identifying whether the T-72 or F-16 belongs to country A, which is an ally, or country B, which is an enemy.

Typical narrow field of view imaging systems for military applications will have FOV in the range 5–0.5°, and Johnson criteria identification ranges from 1 to 10 km.

Image based target classification techniques are being developed with a goal of achieving a reasonable level of ‘classification’ confidence—for example tank rather than cow or car, jet fighter rather than civil airliner—as an aid to clutter discrimination in systems where the operator is automatically alerted to possible targets.

Nonimaging techniques for classification and tracking

Image based automatic classification techniques are augmented by other discriminants such as the trajectory, spectral signature, and temporal signature of the feature. All these methods are used to increase the robustness of infrared search and track (IRST) systems and missile warners. They are also used in automatic video trackers, which are used to keep the optical line of sight of imaging systems pointing at the target as the target and/or the platform manoeuvres. These are now able to determine the position of a target in a video image to within one or a few pixels in the presence of clutter, decoys and target aspect changes.

Identification of a target type is often difficult in automatic systems because methods used by the trained observer are based on combinations of cues which are difficult to impart to automatic systems. Other techniques used to improve automatic target ‘identification’ or provide images usable by human observers at longer ranges include laser-gated imaging and laser based frequency domain methods.

C3.3.3.3 Laser systems

Laser rangefinders

Laser rangefinders are used for weapon aiming systems (particularly armoured vehicle fire control), for target location in surveillance, reconnaissance and artillery observation roles, and for navigation, to determine the user’s distance from a recognizable navigation feature (figure C3.3.4).

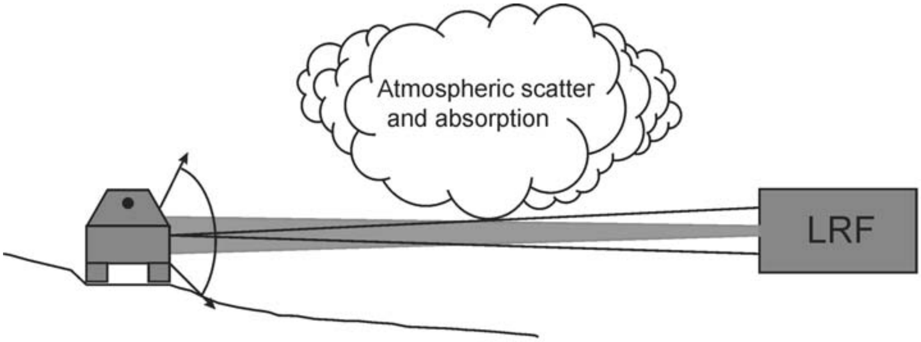
Most of these systems use a relatively high powered short pulse (typically many millijoules in 10–20 ns) and measure the time taken for radiation scattered from the target to return to a sensitive receiver accurately aligned to the transmit beam. They are always associated with direct or indirect sights (figure C3.3.5).

The first military laser rangefinders used flashlamp-pumped Q-switched ruby and Nd:YAG lasers operating at 693 and 1064 nm. They presented a significant eye hazard because these wavelengths are transmitted and focused by the human eye, so relatively low power pulses can create high, potentially damaging, power densities on the retina. Therefore they could not be used freely in training. This in turn made their effective use in combat less certain because operators were not completely familiar with using them. Many laser rangefinders entering service now use the ‘eyesafe’ 1.54 μm wavelength—radiation of this wavelength is absorbed within the eye and does not focus down on the retina, so the maximum permitted exposure is much greater than for shorter wavelengths. With careful balancing of transmitted power and receiver sensitivity, modern ‘eyesafe’ laser rangefinders provide militarily useful range performance and can be used with minimal or no operational restrictions.

Most current military LRFs are one of three basic types:

- Nd:Yag or Nd:glass lasers working at 1.06 μm.
- Raman-shifted or OPO-shifted Nd:YAG producing ‘eyesafe’ output at 1.54 or 1.57 μm.
- Erbium glass producing 1.54 μm output directly.

Most are flashlamp pumped, but diode pumped systems are starting to enter the market. Q-switching techniques include spinning prisms (used in the first military rangefinders and still common), EO Kerr or Pockels cells, and saturable absorbers. CO_2 ‘TEA’ lasers are used in some military


Atmospheric scatter and absorption
LRF

(1) The laser boresight mark is aligned to target.
(2) The laser pulse is fired at the target (typical pulse lengths are 10–20 ns).
    (i) a small percentage is bled off to the receiver to start the clock.
(3) The laser energy is attenuated by scatter and absorption in the atmospheric path.
(4) With a sufficiently narrow beam divergence most or all of the remaining energy in the beam hits the target.
(5) A percentage of the laser energy is scattered by the target:
    (i) most targets can be approximated as lambertian scatterers with a diffuse reflectance of 10–40 %.
(6) The energy collected by the receiver aperture depends on:
    (i) the solid angle subtended by receiver aperture seen from the target.
    (ii) and atmospheric losses in the return path.
(7) The collected energy is focused onto the detector. Receiver sensitivity depends on a number of factors including detector noise, background noise, quantum efficiency and optical transmission.
(8) A laser rangefinder range equation can be derived from this geometry.

Figure C3.3.4. Laser rangefinder principle of operation.

systems, but have not been widely adopted because they tend to be bigger and more expensive than VNIR systems for a given performance.

Detectors are typically silicon PIN diodes or Avalanche photodiodes for 1.06 μm. Eyesafe laser receivers benefit from the enormous progress in telecoms systems at 1.54 μm and use one or other of the

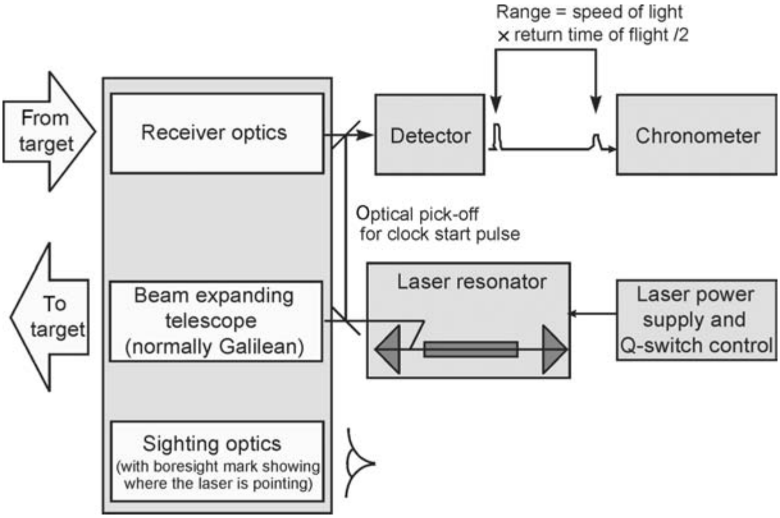

Range = speed of light
× return time of flight /2
From target
Receiver optics
Detector
Chronometer
Optical pick-off
for clock start pulse
To target
Beam expanding telescope (normally Galilean)
Laser resonator
Laser power supply and Q-switch control
Sighting optics
(with boresight mark showing where the laser is pointing)

Figure C3.3.5. Laser rangefinder generic architecture.

GaAs-based compounds. CO_2 systems usually use cadmium mercury telluride detectors cooled to liquid nitrogen temperatures. Typical effective ranges are 5–10 km for battlefield systems, more for high power airborne and ship-borne systems.

Laser rangefinders need to be accurately aligned to the crosswires in the aiming system, otherwise the range obtained will be to the wrong target. They also need to be very robustly engineered to maintain alignment in the severe environment experienced in most military systems.

Laser designators

Laser designators are used to provide the guidance illumination for laser guided ordnance. They are universally Nd:YAG lasers, working at pulse repetition frequencies in the 10–20 Hz region. This is unlikely to change in the foreseeable future because of the massive existing inventory of seekers and designators. Some designators are designed as dual wavelength systems, with a 1.54 μm capability for rangefinding. Most designators have a secondary laser rangefinding capability if only to confirm the correct target is being illuminated.

Land based designators are mostly lightweight, tripod mounted, man portable, used by Forward Air Controllers and Special Forces.

Airborne laser designators are integrated into designator pods for fast jets, or turrets for helicopters or UAVs. These airborne systems typically have a high performance thermal imager, TV camera or both; a video tracker; and high performance stabilization, to allow the system to lock onto a target and maintain a stable line of sight as the aircraft manoeuvres.

Designator lasers operate at a high internal power density and are required to have very stable and repeatable boresight, beam divergence, pulse repetition frequency and pulse energy. They typically use techniques pioneered in the early 1970s in the UK and US: the ‘crossed Porro resonator’ [20], EO Q-switches, polarization output coupling, great attention to choosing optical architectures and components which are inherently stable, careful thermal management, and rigorous contamination control. Rod and slab lasers are used; rod lasers need particular attention to their warm-up characteristics since the optical power of the rod tends to change during laser warm-up. Flashlamps are the commonest pump source, but diode lasers are beginning to be used instead. Diode pumping of Nd:YAG lasers is much more efficient than flashlamp pumping, but more expensive and (at least until recently) with significant thermal management problems of its own.

High energy lasers

After a long period of research and demonstration, high energy lasers are now being used in developmental anti-missile systems and have demonstrated an ability to cause airframe damage. These systems use chemical lasers, where a chemical reaction creates the energy level population inversion required to allow laser action. This mechanism is very efficient compared to most electrical excitation methods, which makes it feasible to achieve the required very high power levels in a militarily useful package. Possible use of space-based lasers is discussed by Rogers [24].

Laser-gated imaging

To achieve longer range in a wider range of ambient lighting and obscuration conditions, a technique known as ‘range-gated active (or laser) imaging’ can be used. A pulsed laser (usually in the visible or near IR) is fired at the target, and a time-gated imaging sensor is used to capture a snapshot of the target using the scattered laser light. By matching the time-gating to the round trip time of flight, the receiver is switched on just before the scattered laser energy arrives back from the target and switched off again just after. Atmospheric backscatter and background clutter can be rejected by the time gating, leaving the target standing out in an otherwise blank image. Such systems can obtain high resolution imagery at long range with small apertures, independent of ambient light and unaffected by line of sight jitter.

This technique was first developed in the late 1960s. It has also been used in a commercially available system for search and rescue in Canada called Albedos [6].

Laser radar

Laser radar systems have been developed for obstacle avoidance. Typically they transmit a rapid sequence of laser pulses over a pre-determined scan pattern. A receiver detects the return energy scattered from the ground and from wires or pylons; timing discrimination allows the latter to be distinguished from the former. Warning cues are displayed to the pilot to allow him to take appropriate avoiding action.

Laser beamriding

Lower power lasers, often near-IR laser diode arrays, are used in laser beamriding systems (see the section on weapon guidance).

C3.3.3.4 Important enabling technologies

As well as lasers and imaging, several other technologies are key to the engineering of successful optoelectronic systems for military applications and other similarly harsh environments.

- *Stabilization*. Line of sight stabilization systems typically use gyros to sense disturbance of the line of sight or the optical chassis, and a servo controller and mechanism to adjust the line of sight to compensate for the disturbance. Performance (measured as achieved jitter level under angular disturbance and linear vibration) varies widely, from sub-micro-radians in spacecraft, through 5–20 μrad in high performance airborne systems, to 50–500 μrad in typical land vehicle systems. Stabilization performance is best understood as an attenuation of the input disturbance, so performance figures must be taken in context with the environment. Harsh vibration and shock environments also impose design constraints on stabilization that tend to degrade the achievable performance [15]. Anti-vibration mounts are often used to give the critical stabilization components a softer ride, prevent linear vibration from exciting structural resonances, and match the angular disturbance spectrum to the control bandwidth of the stabilization servo.
- *Processing*. Modern EO systems are largely software driven. The relentless march of Moore's law means that digital processing is becoming the default solution to most control problems, and embedded computing power is now sufficient for real-time image processing functions such as detector channel equalization, contrast optimization, image enhancement, and target detection, tracking and classification. Many image-processing systems developed during the 1990s used custom silicon devices (application specific integrated circuits or ASICs). At the time of writing, military production volumes are of little interest to ASIC manufacturers. Field programmable gate arrays (FPGAs) and digital signal processors (DSPs) are now widely used in military systems. Mass produced devices with high data throughput capability, they are programmed by the individual developer; so are more flexible and more economical than ASICs for low-volume applications such as military systems.
- *Lens design*. While modular design and re-use are becoming more common, most military systems require custom optical design to satisfy installation constraints and system performance requirements. Particular challenges for the lens designer in military instruments include design

for mechanical robustness, high boresight stability and repeatability, athermalization, and good performance in the presence of bright sources in the FOV (e.g. fires, sunlight). There are numerous references to this subject, notably proceedings of lens design conferences run for example by the SPIE. IR optical designs are substantially different from visual band ones because of IR materials' higher refractive index, wider range of dispersions, and larger thermal refractive index coefficients.

- *Optical materials*. Military requirements have driven development and precise characterization of novel optical materials for many uses within EO systems. These include: IR lenses for various subregions of the spectrum; mechanically strong broad-band windows with good thermal properties; laser and other components with low absorption, low scatter and high laser damage threshold; and lightweight, thermally and dynamically stable mirrors and structures.
- *Optical fabrication*. Diamond turning now allows the reliable and repeatable fabrication of aspheric surfaces, which allow simpler and higher performance IR system designs. 'Binary' or diffractive optical components can also be made with diamond turning; this allows colour correction (in systems where scatter is not a critical design driver) with even fewer optical components. Diamond fly-cutting is widely used to finish flat and cylindrical mirrors. Great strides have been made in the surface finish, accuracy, repeatability and flexibility of these techniques over the last 20 years, driven mainly by military requirements.
- *Thin film coatings*. Thin film coating products and custom engineering capabilities have been developed to satisfy the many and varied demands of military EO systems. These include: low loss laser coatings with high damage threshold and low scatter; efficient wavelength selective beamsplitters and filters; high efficiency coatings for a wide range of spectral regions; multi-spectral coatings simultaneously offering good performance (usually high transmission) over a number of spectral regions; exceptional environmental durability (sand erosion, rain impact, chemical contamination, wide temperature range); all compatible with a range of substrate refractive indices and thermal expansion coefficients.
- *Optomechanical design*. Military optoelectronic systems usually need to be compact, lightweight, and robust to a wide range of temperatures and under shock and vibration; easy to align and maintain, hence modular in construction; and to achieve high line of sight stability. Often they incorporate mechanisms for switching components in and out of the light path, for scanning or steering the line of sight, and for adjusting focus and magnification. Embedded electronics generate heat, which has to be managed without adverse effects on the rest of the system. Delicate optics have to be housed to maintain critical alignment yet survive severe shocks. These create tremendous challenges for mechanical and optical engineers, leading to the discipline of 'optomechanical engineering'. Vukobratovich [29] and Yoder [30] address many of the detailed design issues. Godfrey [16] and Jamieson [17] provide insights into the complex issues that occur at the 'system architecture' level.
- *Simulation and modelling*. Parametric modelling and physically accurate simulation techniques have been developed to support system requirements definition and validation, system trade-offs, detailed design, and validation of the final system design. Accurate techniques have been evolved to model all aspects of EO system performance, including scene illumination, target and background signatures, atmospheric absorption and emission, scintillation induced by atmospheric turbulence, image formation, scatter, line of sight stability, and detection and tracking performance. Such models may be integrated to provide realistic end-to-end simulations displaying a representative image from a system yet to be built, or of a scenario too dangerous or expensive to trial—or both. Such models are quite different from those used in training simulators. Simulator systems provide

an impression sufficient for their purpose of what an observer would see, using approximations and assumptions to simplify computation and to ensure real time performance. EO system models provide a physically correct representation of what an observer or image-processing system would see with a real instrument in a specific scenario.

- *Calibration, test and alignment*. Numerous test techniques have been developed: both to align, integrate and characterize EO systems as 'black boxes' in their own right; and to assist with their integration with other systems, for example platforms and weapon systems. Methods of particular note include those associated with characterizing and correlating the subjective and measurable performance of thermal imaging systems; techniques for aligning multi-waveband sensor systems; and 'auto-boresight' techniques for dynamically maintaining system alignment during operation.

C3.3.4 Environmental factors

We have already alluded to a variety of environmental factors that influence military EO systems. In addition to 'the EO environment', already discussed, these can be summarized as:

- Physical threats: ballistic, rain and sand erosion, air pressure, hydrostatic pressure, own and hostile weapon blast effects.
- Operating environment—shock, vibration, temperature, handling, radio frequency interference.
- Line of sight limitations.

C3.3.4.1 Physical threats

Physical threats to EO systems include:

- Ballistic—for example bullets and shrapnel may hit delicate optics, causing catastrophic damage. System design is a trade-off between maximizing performance (usually requiring a large aperture) and minimizing vulnerability (for example by keeping the aperture small). Many systems are designed with shutters which can be lowered to protect the optical aperture, often with slits which allow continued operation with degraded performance.
- Rain and sand erosion—forward facing optics fitted to fast low-flying aircraft are particularly susceptible to damage from sand erosion, which can strip coatings and pit surfaces, and to rain impact which above a velocity threshold can cause serious sub-surface damage, causing rapid loss of transmission. Coatings such as boron phosphide and diamond-like carbon can be applied to protect surfaces while maintaining good optical transmission.
- Chemical erosion may be caused by seawater and by exposure to fuels, cleaning agents and exhaust efflux. Again, inert coatings can usually be identified to protect while maintaining optical performance.
- Air pressure, hydrostatic pressure, own and hostile weapon blast effects all impose structural loads on external windows which normally require thick windows to maintain structural integrity and careful mounting to avoid excessive local stresses which can initiate structural failure.

C3.3.4.2 Electro-magnetic threats

Enemy systems may use EO or radar sensors to detect and/or degrade EO systems. Notably, submarine periscopes are subject to detection by radar, and reports in the trade press suggest that hostile lasers intended to damage detectors are also seen as a threat. Filters can be fitted to minimize the effect of EO countermeasures but this is often at the expense of the performance of the EO device. Increasingly attempts are being made to develop detectors that operate outside normal threat wavebands.

C3.3.4.3 Platform characteristics

Different platform types have widely different environmental requirements, mission characteristics, and accepted industry quality standards and interfaces. They also imply different user characteristics: education, training, and tolerance of workload, usability, quality, reliability, and technology complexity; and different production volumes, rates, and support and maintenance philosophy.

Notably, equipment on helicopters has to tolerate very strong low frequency vibrational resonances at harmonics of the rotor frequency. Equipment on main battle tanks (MBTs) and submarines must be hardened to operate after ‘survivable hits’ from shells or depth charges, sometimes with accelerations specified up to hundreds of g. Equipment on aircraft must operate in the presence of high linear vibration caused by engines and aerodynamic buffeting, and high angular disturbance caused by the aircraft’s manoeuvres.

EO sensors often have to operate near other systems which generate strong RF fields: for example radars, communication equipment.

C3.3.4.4 Line of sight limitations

With the rare and uncertain exception of atmospheric ducting, free space EO systems cannot see round corners. Strong signals, for example from missile plumes, can be detected via atmospheric scatter when there is no direct line of sight. Otherwise EO systems depend on a clear line of sight from sensor to target.

So terrain geometry strongly influences the use of military EO systems—even to the extent that different countries will specify quite different equipment characteristics, depending on whether they expect to be operating with long open lines of sight, for example desert conditions, or short lines of sight in forest, urban or rolling rural conditions. It is reported that historically most tank engagements occur at less than 1 km, although most modern MBTs are designed to work at up to 3 km. In the 1991 Gulf War, line of sight ranges were often very long, and armour engagements at ranges of up to 5 km were reported. Tactics must be adapted to make best use of the characteristics of available equipment in different terrain.

At sea, the horizon and weather limit lines of sight. In the air, cloud and terrain normally limit the line of sight at low altitude. At medium and high altitude (above 5000 m) and in space, air-to-air lines of sight are very long, limited only by the earth’s curvature. At high altitude, the atmosphere is thinner and atmospheric absorption becomes less significant. Air to ground visibility is (obviously) affected by cloud cover.

C3.3.5 Roles and examples of military EO

The following sections discuss the main roles in which EO systems are used by the military, under the six principle categories of:

- reconnaissance;
- surveillance and target acquisition (S&TA);

- target engagement;
- self-defence;
- navigation and piloting; and
- training.

The final section deals with some less common systems that do not fit into any of these categories.

Each section discusses the role, gives examples of systems, and discusses some of the underlying design issues.

C3.3.5.1 Reconnaissance

> "Time spent in reconnaissance is seldom wasted"—"Military Maxims"
> quoted by Mitchell [31].

Reconnaissance is about providing an overall picture of the area of operations, notably where the enemy is and what he is doing. It involves observation and orientation and supports decision-making. There is an increased emphasis on using strategic reconnaissance for tracking and targeting as well. This will require improved real time links between reconnaissance and combat units.

Reconnaissance is performed at all levels, from strategic to tactical. At a strategic level, the area of interest may be worldwide and decision timelines may be in the order of hours, days or weeks. At lower levels the area of interest becomes smaller but the timelines become shorter. At the lowest tactical level the soldier uses his eyes or binoculars, and would dearly love to 'see over the hill'.

Reconnaissance platforms include satellites, manned aircraft, land vehicles and foot soldiers. Unmanned air vehicles (UAVs) are increasingly being used or proposed for the full spectrum of reconnaissance. The US 'Global Hawk' programme is a strategic reconnaissance UAV. Concepts are being proposed for micro-UAVs small enough to be carried by an individual soldier. This will finally make the 'see over the hill' dream a reality for even the lowest level tactical commander.

The human eye has always been used for reconnaissance and always will be. For most of the 20th century wet film was the prime recording medium for aircraft.

The advent of satellite reconnaissance forced the development of solid state image sensors, which now deliver astonishing resolution from large aperture satellite systems in low earth orbit. Fixed wing aircraft are fitted with podded or built-in linescan reconnaissance systems using solid state linear array imagers working in the VNIR, and IR linescan systems working in the 8–12 μm band. Two-dimensional staring arrays with stabilized step-stare mirrors are used for long range stand-off reconnaissance in the VNIR and 3–5 μm band. Improvements in sensitivity are allowing VNIR sensors to be used at progressively lower light levels with progressively shorter integration times.

The key performance goals in reconnaissance systems are to maximize area coverage, resolution and stand-off range of the platform. The ideal reconnaissance EO system would have high resolution, short image capture time, and wide FOV.

Reconnaissance systems

Airborne fixed wing reconnaissance systems are produced by a number of companies in the US, UK, France and other countries. A typical fixed wing low level reconnaissance system is the Vinten Vicon product family. Based on a modular architecture and normally podded, these systems (figure C3.3.6) are used by about 20 air forces and are being configured for UAV applications.

Most armies maintain ground and helicopter units with a reconnaissance role. These generally operate covertly when possible, behind enemy lines, and offer the benefits of having a human observer

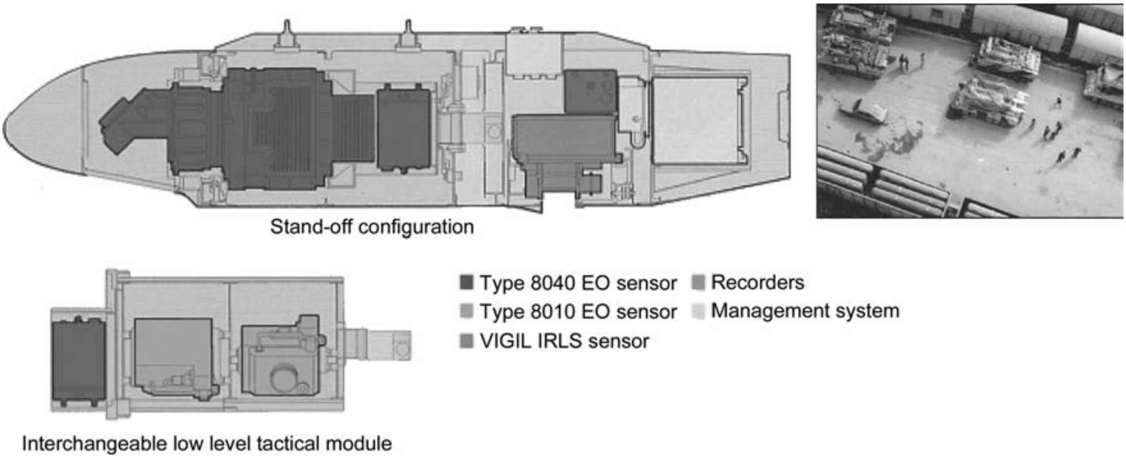


Figure C3.3.6. Typical reconnaissance pod—reconfigurable for low level and stand-off reconnaissance.

able to interpret and prioritize the information, while being relatively immune to cloud cover which can block satellite and UAV operations. Army reconnaissance units are normally equipped with surveillance systems with varying levels of performance, generally aiming to be able to observe over a reasonable area (2–10 km) from a 'hide'. Different armies have widely differing philosophies on how to do this difficult task, depending in part on the terrain they expect to operate in, leading to a remarkable diversity of equipment.

Reconnaissance using manned aircraft or land vehicles is not always politically or militarily practical; satellites and UAVs are becoming widely used for reconnaissance.

UAV reconnaissance systems typically use either or both of linescan sensors derived from aircraft systems, or surveillance turrets derived from helicopter systems. The surveillance turrets typically contain thermal imaging and TV sensors with multiple fields of view, and can be given an area search capability by scanning the whole turret provided the imager integration time is short. The imagery is data-linked to a ground station in real time, and can also be recorded on board for subsequent recovery and exploitation.

Some US 'Predator' UAVs had at the time of writing been given an 'armed reconnaissance' capability with a laser designator coupled with Hellfire laser guided missiles, allowing the UAV to be directed to attack a target observed by the operator at the ground station.

Information about space based reconnaissance systems is harder to come by. Chaisson refers to KH-11 Keyhole satellites used by the US. News reports during recent conflicts have emphasized the use of satellite reconnaissance by the US, and the possible military use of commercially available satellite imagery from many operators throughout the world. US government officials' statements which can be found on the web emphasize the priority attached to maintaining a 'technological advantage' in the capture and exploitation of information using satellites.

Resolution of reconnaissance systems

Popular fiction suggests that reconnaissance systems can read car number plates from space. To read a number plate the system would need to resolve about 1–2 cm. We can test this assertion with a simple calculation based on the Rayleigh criterion for optical resolution:

$$\theta = 1.22\lambda/d$$

where θ is the minimum angular separation between two resolvable points, λ is the wavelength and d is the diameter of the aperture. The reciprocal of this angle corresponds to the cut-off frequency (typically quoted in cycles per milli-radian) of the modulation transfer function of a diffraction-limited imaging system.

A 600 mm aperture telescope operating at a wavelength of 500 nm would have a diffraction limited optical resolution of 1 μrad. This would allow an imaging resolution of 1 m at a height of 1000 km in the absence of image motion and atmospheric degradation. Equivalent resolution for a system working at 10 μm in the thermal IR would be 20 m. This example illustrates the benefit of using the VNIR band where target signatures allow.

Some US reconnaissance satellites are similar in general characteristics to the Hubble Space Telescope [5], with its 8 ft (2.4 m) mirror. The International Space Station is approximately 400 km above the earth, and orbits at 27,600 km hr^{-1}. (NASA web site). A Hubble-class telescope in this orbit, operating in the visible spectrum at 500 nm, would resolve 0.1 m, or 4 in, on the ground, if its line of sight was correctly stabilized on the earth's surface, and if the effects of atmospheric turbulence were overcome.

C3.3.5.2 Surveillance and target acquisition (S&TA)

Reconnaissance imagery is very data-intensive. It tends to be analysed by interpreters who summarize the information and provide it to senior commanders. It is as yet seldom available in real time to troops on the ground, because the bandwidth required to disseminate raw imagery is very high and seldom available.

Tactical real time targeting and situation awareness is provided by wide or medium FOV imaging sensors which can be panned or scanned to cover a wide field of regard. Typical solutions are surveillance turrets, typically carrying a thermal imager, TV camera and eyesafe laser rangefinder.

Similar sensor suites are integrated into the turrets of many military vehicles for surveillance, target acquisition and fire control, mounted on pan and tilt heads for border and base security, and on vehicles with elevating masts for mobile patrols. Compact and increasingly integrated systems with similar functionality are used by forward observers, usually on manually operated tripods, and will shortly be small and cheap enough for use by individual soldiers, albeit with scaled down performance. Such systems are sometimes integrated with navigation systems to allow the target grid reference to be calculated.

System example: surveillance turrets

Turret systems are used on helicopters, patrol aircraft and some land vehicle applications.

They are typically stabilized, to a level appropriate to the resolution of the sensor payload. They are used for general purpose surveillance, tracking, and targeting, for rescue, and as navigation aids to the crew. This class of product has received widespread publicity on western television because TV programmes such as 'police, camera, action' make extensive use of the often spectacular video sequences obtained by police helicopters during surveillance and pursuit operations, mostly car chases.

They usually look like a ball, with an azimuth gimbal which allows the whole ball to rotate, and an elevation gimbal which allows the payload within the ball to elevate. Some have a third gimbal, which allows the system to compensate for roll and to keep the image upright while tracking through the nadir. Others have a second pair of gimbals for fine stabilization of the payload. This separates the functions of steering, which is assigned to the outer pair which also has to cope with friction from the environmental seals, from that of stabilization, assigned to the inner pair which only have to work over a limited angle assisted by the intrinsic inertia of the payload.

Large, high performance turrets with extremely good stabilization (from 20 μrad down to a few microradians) are used for long range stand-off surveillance, usually with large aperture TV cameras.

Small and agile turrets fitted with thermal imagers and/or intensified CCD cameras are used as visually coupled flying aids. The line of sight is steered to follow signals from a head tracking system which measures where the pilot is looking; the image is overlaid on the scene by the pilot's helmet mounted display. Small agile turrets are also used for active countermeasures systems.

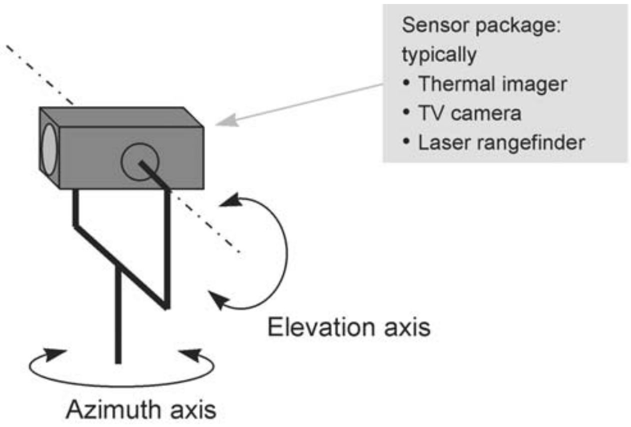
Similar systems with appropriately customized or simplified packaging, pointing and stabilization are used for other surveillance and fire control applications; for example border surveillance, and ship-borne surveillance and fire control. On ships they are generally referred to as 'directors'.

The generic principle of operation of surveillance turrets and 'directors' is shown in figure C3.3.7.

Infra-red search and track

Most S&TA sensors are designed primarily to provide images to operators who are responsible for detecting and classifying targets. IRST systems are a special case of S&TA systems designed to detect and track certain classes of targets automatically without operator intervention. Their main application is against aircraft, which are relatively easily detected because of their high contrast against cold sky backgrounds and their distinctive motion relative to ground clutter. Similar principles can be used from space to detect ballistic missiles.

IRST systems give their users big tactical advantages. They allow them to detect and track aircraft without revealing their presence with radar emissions; they allow more accurate angle tracking of targets than is possible with radar; and most stealth technologies are not as effective against IRST systems as they are against radar. Range to target can be determined, albeit not as accurately as by radar, using kinematic or triangulation techniques. Highly effective as stand-alone systems, IRST systems are even more valuable when integrated into a sensor network including radar, since each sensor to a considerable extent complements the other.



Notes:
Typically, slip rings are used for electrical connection across the azimuth bearing to allow unrestricted nx360 degree rotation.
'3-axis' systems have a third rotation axis orthogonal to the other two to provide roll stabilization.
'4-axis' systems have an outer coarse set of gimbals carrying an environmental cover, and a fine inner pair for precise stabilization.
'5-axis' systems have a 2-axis outer set of gimbals for environmental protection and a 3-axis inner set for stabilization and derotation.

Figure C3.3.7. Principles of surveillance turret operation.

IRST system concepts date back to the 1930s, when R.V. Jones demonstrated aircraft detection at operationally useful ranges using cooled detectors, probably PbS. Radar was selected for further development at the time because of its all weather capability.

First generation airborne IRST systems were fitted to the Mig 29, SU27, (both 3–5 μm) and F-14 (8–12 μm band). Their design varied; at the simplest they were conceptually similar to missile seekers with programmed scan patterns. Second generation systems fitted to the Eurofighter Typhoon and the Dassault Rafale use high performance detector arrays allowing beyond visual range detection, and are in a real sense 'passive radars'.

Some ships have been fitted with scanned IRST systems to fill the radar gap in detection of sea skimming missiles. Successful production examples include the Dutch Sirius and IRScan systems, using first and second generation long linear array technology respectively, and the French Vampir. The US, Canada and the UK have been developing naval IRST technology for many years, but at the time of writing had developed no product.

Of land-based systems, the Thales air defence alerting device (ADAD) has been notably successful, achieving large orders from the British Army. Working in a static ground clutter environment it gives fully automatic passive alerting of approaching air threats.

Naval and land based systems work in a very complex clutter environment and for best performance require good training and a clear doctrine for systems integration and deployment. As technology develops it will be possible to improve the efficiency and robustness of IRST clutter discrimination, widening the applicability and operational flexibility of these systems.

System example: submarine periscopes

Submarine periscopes provide a view of the outside world when submarines are on the surface and at periscope depth. They are used for general navigation and watch-keeping, surveillance, threat and target detection, and to identify and range to potential targets.

Traditional periscopes are all-glass, relaying a magnified view of the scene to the control room, typically 10 m below the line of sight. This is a demanding optical design problem, with only a small number of companies in the world capable of providing effective systems.

Conventional periscopes now have EO sensors integrated into them. For example, the Thales CK038 product family offers thermal imaging, intensified CCD, colour TV and still camera channels as well as 2 or 3 direct optical fields of view. The indirect sensors can be displayed on a remote console, and 'bearing cuts' and 'range cuts' provided to the combat system. A variety of antennae for RF sensors can also be carried.

Nonhull-penetrating masts are the next technology step. Thales is supplying 'optronic masts' for the new Royal Navy ASTUTE class submarines which will be the first in the world to have no direct optical path from inside the hull to the outside world. This innovation reduces the number of holes in the pressure hull and removes constraints on the position of the control room relative to the submarine's 'fin'. Exceptionally high quality systems with excellent demonstrated performance, intuitive man machine interface, and built-in redundancy are required to give users confidence in this approach.

C3.3.5.3 Engagement

Weapon aiming

Weapon aiming involves:

- detecting the target or acquiring it after being cued on by another sensor;

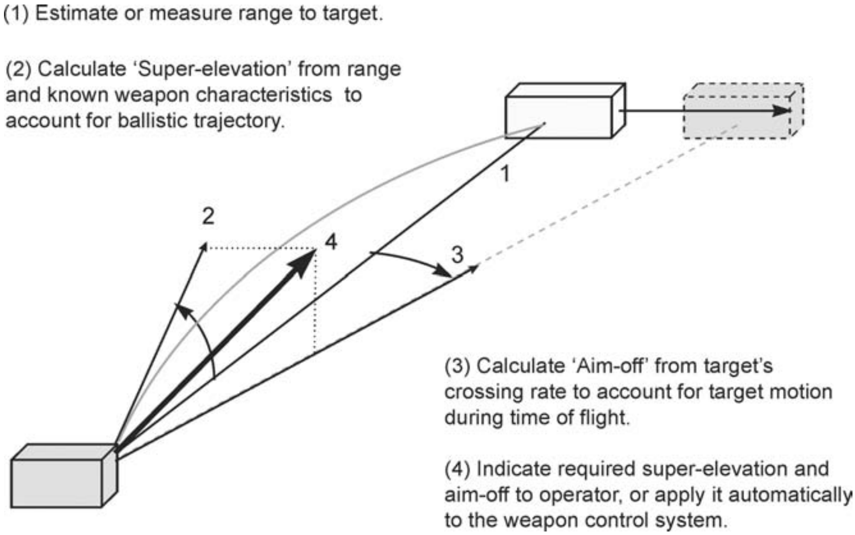
- determining the information required to launch the weapon—normally target direction, range and crossing rate, and often also range rate; and
- displaying the results of any fire control computation to the operator to allow him to point the weapon in the right direction for launch.

Typical weapon aiming accuracies for armoured vehicle gun systems are in the 50–500 μrad range (10–100 arc s). Shocks in the range 40–400 g may be generated at or near the sensor by the weapon launch. A key part of the skill in weapon aiming system design is to package delicate optical instruments in such a way that they will retain the required accuracy, yet survive the extreme shock and vibration on weapon platforms.

System error contributions come not only from within the EO system, but:

- from the weapon system itself—for example barrel wear and bending under thermal gradients, geometric imperfections and play in linkage mechanisms, and tilt of the weapon platform;
- from the environment—cross-winds and air pressure variations;
- from dynamic effects within the weapon platform—vibration, flexing and servo latencies; and
- from target motion—the target may move significantly and unpredictably during the time of flight of the projectile.

Modern high performance fire control systems measure these contributions and take them into account in the ballistic calculation, minimizing their effect and achieving high overall system accuracy. But conventional gun systems cannot adjust the course of the projectile after it leaves the barrel. So they cannot compensate for target acceleration, and lose accuracy at longer ranges as the projectile slows down and becomes less stable in flight (figure C3.3.8).



(1) Estimate or measure range to target.

(2) Calculate 'Super-elevation' from range and known weapon characteristics to account for ballistic trajectory.

(3) Calculate 'Aim-off' from target's crossing rate to account for target motion during time of flight.

(4) Indicate required super-elevation and aim-off to operator, or apply it automatically to the weapon control system.

Figure C3.3.8. Generic weapon aiming system.

Guidance

Aiming systems for unguided projectiles need to point the weapon accurately at launch and to anticipate the target's motion between launch and impact. These requirements for precision are reduced if the target can be tracked and course corrections made during the flight of the missile. This can be achieved either by tracking the target from the launcher or by homing, either semi- or fully autonomous.

Common techniques used are:

Command guidance

- command to line of sight
- beamriding

Smart/terminal guidance

- laser designation
- autonomous homing

Command guidance

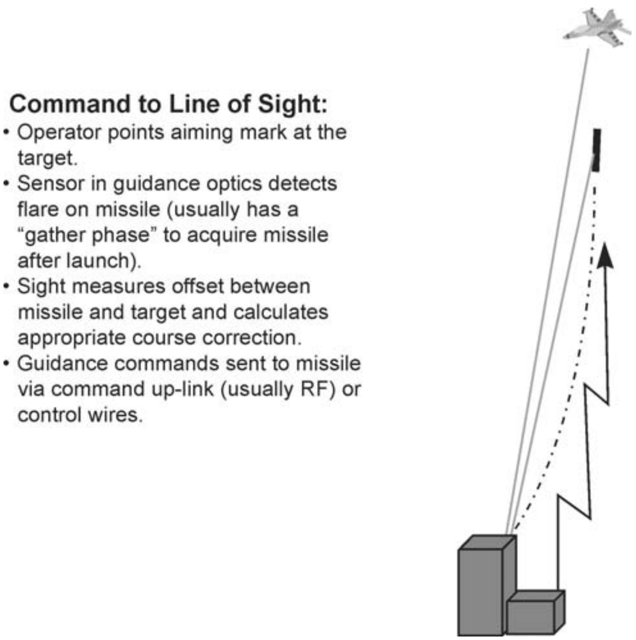
In command to line of sight (CLOS) systems, both target and missile are tracked. In many systems, the operator keep a crosswire on the target and a sensor in the sight tracks a flare on the tail of the missile. More modern systems track both target and missile automatically. A guidance system measures the error, and an RF, EO or wire link transmits control signals to the missile. These systems often use a rangefinder to measure range to the target, and keep the missile off the line of sight to the target until near the expected impact time to prevent the missile or its plume obscuring the target (figure C3.3.9).

In a beamriding system, the operator keeps the crosswires in the sight on the target, and a spatially and temporally modulated laser beam is projected along and around this line of sight. A receiver in the tail of the projectile detects the signal; signal processing decodes the modulation and generates the appropriate guidance correction. This method relies on the operator accurately tracking the target throughout the engagement. It is used in a number of successful anti-aircraft missile systems (including the Swedish RBS-70 and the British Starstreak) and some anti-tank missile systems. Most Western beam riding systems use diode lasers, a small number use CO_2 lasers. Power levels are much lower than for other military laser systems because it is a one-way co-operative system (figure C3.3.10).

Homing

Laser designation (figure C3.3.11) involves illuminating the target with a pulsed laser which is accurately boresighted to a crosswire in the sighting system. Normally pulsed Nd:YAG lasers are used, associated with a silicon quadrant detector in the seeker head. The seeker, or an optical element within it, is gimballed. A control system moves the seeker line of sight to keep the laser spot on or near the centre of the quadrant detector. The missile in turn flies to follow the seeker line of sight. This system, pioneered by the USA during the Vietnam war, is regarded as the most accurate precision guided system and is in extremely widespread use for anti-tank missiles, air-launched missiles and a wide range of bombs. Laser guided artillery shells have been developed but are not in widespread use.

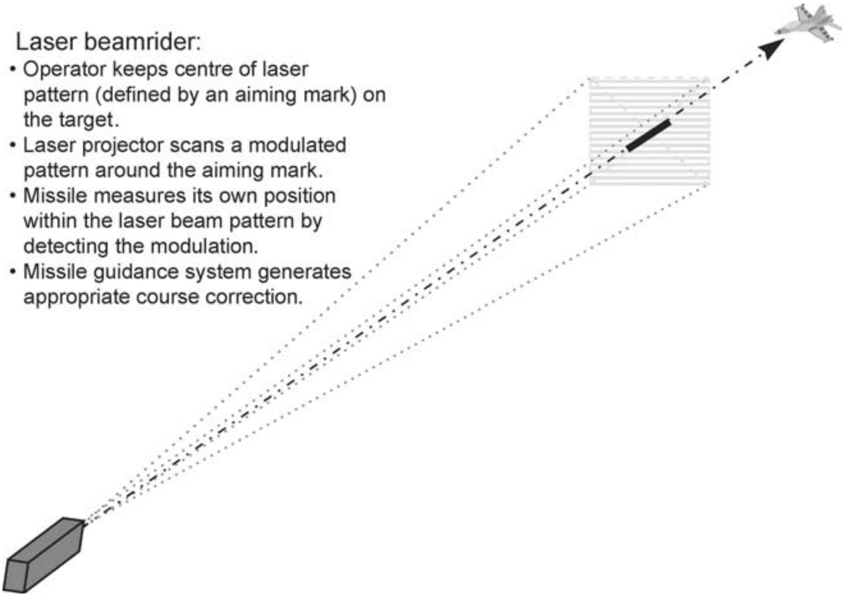
Laser seekers are referred to as '*semi-active* homing' systems—*active* because an active transmitter (the laser) is used to illuminate the target, *semi* because the transmitter is not co-located with the seeker.



Command to Line of Sight:

- Operator points aiming mark at the target.
- Sensor in guidance optics detects flare on missile (usually has a “gather phase” to acquire missile after launch).
- Sight measures offset between missile and target and calculates appropriate course correction.
- Guidance commands sent to missile via command up-link (usually RF) or control wires.

Figure C3.3.9. Guidance principles: CLOS.



Laser beamrider:

- Operator keeps centre of laser pattern (defined by an aiming mark) on the target.
- Laser projector scans a modulated pattern around the aiming mark.
- Missile measures its own position within the laser beam pattern by detecting the modulation.
- Missile guidance system generates appropriate course correction.

Figure C3.3.10. Guidance principles: beamriding.

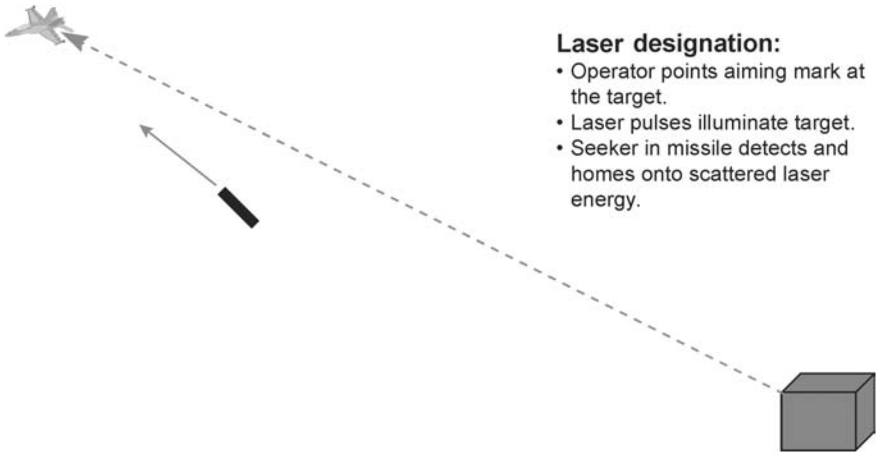

Laser designation:
• Operator points aiming mark at the target.
• Laser pulses illuminate target.
• Seeker in missile detects and homes onto scattered laser energy.

Figure C3.3.11. Laser designation.

Passive missile seekers detect and home onto contrast between natural radiation emitted or reflected by the target and the background. They can be broadly categorized as:

- ‘hot-spot detectors’; and
- ‘imaging’ seekers.

Hot spot detectors are simple, are easily implemented as all-analogue systems, and are effective in low clutter environments such as an aircraft against sky background. Anti-aircraft systems have evolved through several generations to improve sensitivity and countermeasure resistance. The simplest require a strong signal, for example from a hot exhaust pipe, so can only attack from behind, and are relatively easily decoyed or jammed. The more sophisticated are sensitive enough to home on black body emission from an airframe, so are ‘all-aspect’, and employ various anti-decoy techniques, some of them dependent on digital processing.

IR hot-spot seeking surface to air and air to air missiles have accounted for a huge majority of all air combat losses since the 1950s. The best known to western readers are the American Sidewinder family of AAMs, which has evolved incrementally through numerous versions since the 1950s [19], and the Stinger series of SAMs, which evolved from the 1960’s Red-eye, and has itself progressed through at least three major versions using two different detection principles. Other nations including France, USSR, UK, China and Brazil have developed indigenous IR SAMs and AAMs, some using completely independent seeker principles, others based closely on the successful American designs.

Most anti-aircraft seekers work in the 1–2.5 or 3–5 μm bands. Early Sidewinders used uncooled PbS detectors working in the 1 μm region with a glass dome. These systems were easy to engineer using available materials and technologies, but because of the low sensitivity of the uncooled seeker, they needed to look up the hot jet pipe. In the late 1950s and 1960s sensitivity was improved to allow wider aspect but still tail-on engagements. The PbS detectors were cooled with Peltier coolers, which could be operated continually while the missile was on the launch aircraft, or with high pressure argon gas, which achieved a lower temperature for a limited period after a very fast cooldown. Later versions were more sensitive still. They used InSb detectors operating in the 3–5 μm band, again cooled with argon, to achieve an all-aspect capability by exploiting the lower temperature emission from skin heating for frontal engagements. These systems took longer to develop because they required a whole range of technology developments in optical and detector materials.

Four different optical principles are commonly used in hot-spot seekers to generate a guidance signal from the target's radiation. They are 'AM reticle', 'FM reticle', 'rosette scan' and 'cruciform scan'. In reticle scanned systems, the detector covers the entire instantaneous FOV of the seeker, and a chopper, or 'reticle', modulates the FOV so that small hot areas within the image are preferentially modulated [18]. The phase of the modulation indicates the polar angle, while the amplitude or frequency indicates the off-boresight angle. In AM reticle systems, the reticle is spun to do the chopping, while in FM systems an optical element (or the entire telescope) is spun to nutate the image on a stationary reticle. In rosette and cruciform scanning, the detector is smaller than and is scanned over the FOV, normally by spinning an optical element.

Hot spot seekers work well in situations where the target is unambiguously differentiated from the background by its intensity and basic dynamic characteristics. Where this is not the case, imaging seekers are necessary to allow more complex processing to distinguish target from background. This more complex situation applies to most anti-surface engagements and to more extreme anti-air scenarios, for example those involving ground clutter, complex countermeasures and low signature targets. Modern imaging seekers can be visualized as any conceivable imaging system with a video tracker processing the output to extract the target position from the image. A well-known and widely used example of an imaging seeker system is the imaging IR variant of the Maverick family [25].

In almost all seeker systems, the optical line of sight is pointed at the target using the tracker output, and the missile follows the seeker according to a control law. This may vary from a simple pursuit trajectory (the missile flies straight towards the target all the time by attempting to null the line of sight angle) through an intercept course (the missile points to the predicted intercept position by attempting to null the line of sight rate) to a more complex pre-programmed trajectory designed to optimize flight characteristics, foil enemy countermeasures or both. Seekers generally have a small FOV which can be steered over a large angle. Typical generic seeker parameters are:

Instantaneous detector FOV	0.4–4°
Seeker FOV	2–4°
Seeker field of regard	40°

Directed energy

Lasers are used or proposed for a number of directed energy applications.

- Lasers and arc lamps can be used to dazzle or confuse hostile sensors, particularly missile seekers as part of directed IR countermeasure systems.
- Lasers may be used for sensor damage—there are reports of lasers being used to damage sensors in reconnaissance satellites trying to observe strategic installation.
- Large chemical lasers are now successfully integrated with EO pointing and tracking systems, and these systems have demonstrated an ability to destroy incoming missiles at long ranges: for example the Airborne Laser Laboratory and THEL programmes.

For obvious reasons there is little open information in this area! Suffice it to say that complex technology integration and detailed analysis, simulation and trials are required to develop effective systems.

Damage assessment

It is almost as important to the military to know when an engagement has been successful as to carry out the engagement in the first place. Battle damage assessment (BDA) is essential to know whether to continue attacking a target or to switch to another one. There is often a requirement for EO sensors to continue monitoring the target during and after the engagement and to provide imagery or other data which allows a kill assessment to be made. One aspect of this requirement is for weapon aiming sensors to be able to see through the various EO phenomena produced by the weapon itself.

System example: vehicle sights

Vehicle sights vary widely in complexity, from simple direct view sights to stabilized multi-sensor systems. Figure C3.3.12 illustrates the conceptual layout of a typical high performance tank for control sight with thermal imaging, laser, TV and visual sighting functions.

There is now a trend to lower cost and smaller size, hence wider deployment, as imager technology improves and cheapens. A good example of this evolution is shown by a series of products produced by Thales for the British Army.

- The fire control sensor suite for the Challenger 1 tank in the mid 1980s used thermal imager based on the then-new TICM 2 module, mounted in a barbette on the side of the tank's turret, and elevated using a servo system to follow the gun. The Nd:YAG tank laser sight (TLS) originally developed for the Chieftain was mounted on a cradle at the gunner's station and linked mechanically to the gun.
- The Challenger 2 system used essentially the same thermal imager, mounted on the gun barrel to simplify the system and assure best possible boresight to the gun; and the TLS was replaced by a high performance visual gunner's sight with integrated laser rangefinder and 2-axis stabilization.
- After the Gulf War, the decision was taken to replace the image intensified sights on the Warrior infantry fighting vehicles with a thermal imager. A low cost second generation thermal imager provides almost as good performance as the TICM 2 system on Challenger in a much smaller and cheaper package; the laser rangefinder is eyesafe; a TV camera is integrated into the commander's sight and displayed on a flat panel display which also acts as a terminal for the battle management system; a simple low cost stabilization system provides observation on the move at a fraction of the cost of previous systems; and the whole package is much smaller, lighter and cheaper than the older high performance systems.

C3.3.5.4 Self-defence

Threat warning

EO techniques are used for missile and laser threat warning. Missile warning systems depend on detecting the EO emission of the missile motor. Laser warning depends on detecting the direct or scattered energy from the threat laser.

There are two fundamental problems in the design and use of warning systems. The first is to ensure that the detection probability is high and the false alarm rate low, both due to natural phenomena and to enemy systems which may seek to exploit known characteristics of the warning system. A warner that gives too many false alarms will be switched off by the system operators. The second is to link the warning to useful countermeasures against the threat. A warning system is useful only if it increases the probability of surviving the attack and/or successful retaliation.

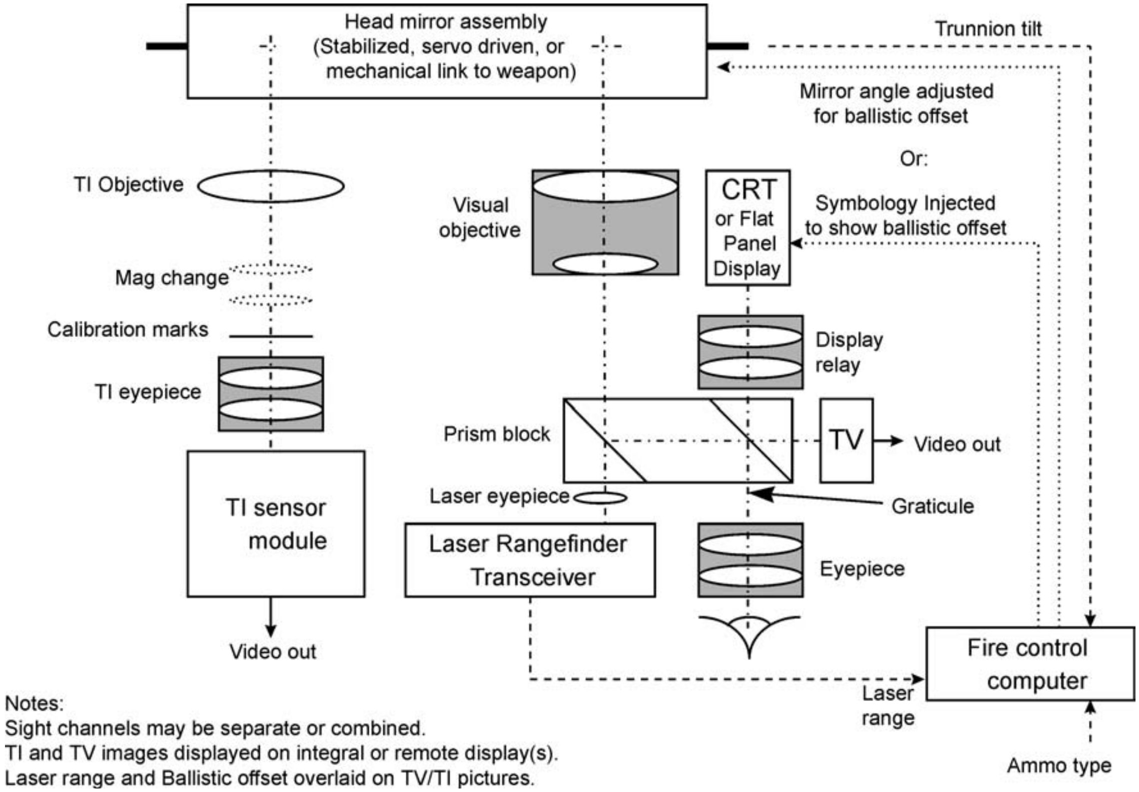

Head mirror assembly
(Stabilized, servo driven, or
mechanical link to weapon)
Trunnion tilt
Mirror angle adjusted
for ballistic offset
Or:
Symbology Injected
to show ballistic offset
TI Objective
Mag change
Calibration marks
TI eyepiece
TI sensor
module
Video out
Visual
objective
CRT
or Flat
Panel
Display
Display
relay
Prism block
TV
Video out
Laser eyepiece
Graticule
Laser Rangefinder
Transceiver
Eyepiece
Fire control
computer
Laser
range
Ammo type
Notes:
Sight channels may be separate or combined.
TI and TV images displayed on integral or remote display(s).
Laser range and Ballistic offset overlaid on TV/TI pictures.

Figure C3.3.12. Schematic of vehicle weapon sight.

These issues have exercised technologists since the 1970s at least. Current in-service missile warning systems are usually associated with active IR countermeasure systems on aircraft, using directional energy (lamps or lasers) or flares or both, and work either in the UV or in the mid-IR. Because of the very short engagement timelines, the linkage between the warning and countermeasure systems is usually automatic. UV warners operating near ground level benefit from a very low background level, since the atmosphere absorbs solar UV before it reaches the earth's surface. By the same token, signal levels are low and the detection range limited, though still operationally useful. In contrast, there is ample signal in the 3–5 μm band, but high natural and man-made clutter because of sun glint, fires and other hot sources. A number of spatial, spectral and temporal techniques are used to classify detections and reject spurious ones before declaring a threat.

Laser warning systems also use a variety of discrimination techniques, including temporal filtering and coherence discrimination, to reject false alarms. Some systems provide accurate direction of arrival information. Deployment of laser warners has been limited because of the 'so what?' issue — if the laser warning was going to be followed within a few seconds by a high velocity shell, what good was the warning? A number of factors are changing this. Improved communication between military vehicles as a result of digitization initiatives will allow threat warning information to be shared quickly between threatened platforms. Until recently, laser rangefinders were normally Nd:YAG, operating at 1.06 μm and therefore a serious eye hazard, and their use was seen as an aggressive act similar to shooting with a gun; but eyesafe lasers will be used much more freely.

Threat suppression

Surveillance, target acquisition and weapon aiming systems aim to help their users by allowing them to detect and engage the enemy. Advantage goes to the side which can complete the OODA loop faster and more consistently. If EO sensors help their user to win, it clearly makes sense to try to reduce their effectiveness, to prevent or delay detection, to make classification more difficult or impossible, to deny the enemy accurate tracking information. The four primary methods for this are camouflage, decoys, jamming, and obscurants.

Camouflage seeks if possible to make the platform look like the background or an innocent natural feature. In World War 2, it was sufficient to break up the visual outline of the platform. With the proliferation of sensors in many different EO bands, camouflage will ideally work simultaneously and consistently in the visible, and near, mid and far IR bands. This is obviously much more difficult, and means that multi-band systems are likely to be able to find even a camouflaged target in at least one of their operating bands.

Decoys seek to confuse the enemy sensor by presenting false targets, and either

(1) overloading the system so that the real target is not noticed among the mass of false ones, or

(2) providing a more attractive target which is tracked in preference to the real one.

Again as sensor systems become more sophisticated, decoys need to be similar to the target in more dimensions simultaneously than used to be the case.

Jamming seeks to overload the sensor so that it is incapable of operating correctly, degrade its sensitivity so that it loses its target, or generate artefacts within the sensor's processing which have the characteristics of real target signals, effectively producing virtual decoys within the sensor processing. A bizarre early attempt to do this (in WW2) is described in [9].

Obscurants, most familiarly smoke, seek to block the optical path between the sensor and its target. The principal mechanism by which most smokes works is scatter. Since scatter is wavelength dependent, with shorter wavelengths being scattered more strongly, most smokes which are opaque in the visible are much less effective, or even completely ineffective, against long-wave IR sensors. This explains the overwhelming preponderance of 8–12 μm systems in NATO armies in spite of the lower volume, aperture and price of 3–5 μm systems of otherwise similar performance.

System example: self-protection and countermeasures systems

Self-protection and countermeasures systems use EO sensors to detect missile launch or laser illumination, and trigger smoke, decoys, or disruptive modulation using wide angle or directed jamming signals from lasers or arc-lamps. This is a sensitive and rapidly evolving area. For a comprehensive unclassified discussion of the subject the reader is referred to the SPIE IR and Electro-Optics Handbook:

- Laser warning (IR&EOSH vol 7 chapter 1.7)
- Missile warning (IR&EOSH vol 7 chapter 1.5)
- Signature suppression (IR&EOSH vol 7 chapter 3)
- Active IRCM (IR&EOSH vol 7 chapter 3)
- Decoys (IR&EOSH vol 7 chapter 4)

- Optical and sensor protection (IR&EOSH vol 7 chapter 5)
- Smoke (IR&EOSH vol 7 chapter 6)

C3.3.5.5 Navigation and piloting

There are three main categories of navigation and piloting applications for EO sensors: helping the pilot or driver see where he is going, particularly at night; helping to keep track of where the vehicle is relative to the earth by terrain or contour matching; and navigating by the stars.

Imaging: flying/driving aids

Thermal imagers and image intensified night vision allow military vehicles to move or fly without active emission, such as lights or terrain following radar.

Fixed wide FOV thermal imagers and intensifiers are used as flying aids or drivers' viewers in aircraft and vehicles. In aircraft the thermal image is often overlaid with accurate registration onto the external scene using a head up display (HUD).

Head mounted NVG, based on VNIR image intensifiers, are also used, usually as an alternative. But both sensor technologies have advantages and disadvantages. VNIR gives a good and intuitively familiar view of the terrain and horizon under all ambient light conditions down to 'overcast starlight'. Thermal imaging on the other hand is better for detecting most military targets (which are usually hot) and many types of man-made features, and provides a means of seeing the terrain when there is no usable VNIR radiation, as occurred during the Gulf War. The RAF has developed integrated system and procedural solutions for using NVG and head-up thermal imaging together. This gives the benefits of both sensor technologies.

A number of helicopter systems and a few modern fast jets use a head steered FLIR coupled to a helmet mounted display (HMD) to provide all-round thermal imaging slaved to the pilot's line of sight. These systems require low latency in the image, the monitoring of pilot's head position, and control of the thermal imager line of sight. Otherwise the image will not appear stable relative to the outside world; objectionable swimming effects may develop in the displayed image which make it difficult for the operators to use the system for any length of time.

There are significant benefits in these applications in sensor fusion, aiming to provide the best of both thermal and VNIR imagery to the pilot. Fully automatic processing is required to reduce pilot workload. Advanced technology demonstrations in this area look promising. The cost of the processing required is coming down rapidly and such systems should soon be affordable for the more high value applications. This illustrates the way that EO system design is now constrained more by data-handling and bandwidth than by the basic sensor technologies.

Terrain matching

Cruise missiles use a number of techniques to navigate very precisely over the terrain. Conventional inertial navigation techniques do not have the required accuracy, having a drift of typically 1 nautical mile per hour. While the advent of global positioning system (GPS) may have reduced the need for other methods, video scene matching and laser altimeter techniques have both been proposed to allow missiles to compare the terrain they are flying over with a stored three-dimensional terrain map of the planned route.

Stabilized EO systems such as surveillance turrets and laser designator pods fitted to an aircraft can measure the aircraft's position relative to an identifiable feature on the ground. If the feature's co-ordinates are known, this information can be used to update the aircraft navigation system, allowing the accumulated drift in the estimated position to be nulled out.

Star tracking

The US and Soviet sea-launched ballistic missile programmes in the 1950s and 1960s wrestled with the problem of hitting a fixed point on earth several thousand miles away from a moving submarine. Part of the solution was an incredibly accurate gyro system on the submarine itself. Another part was the use of star trackers in the ballistic missiles. Star trackers are commonly used in satellites to provide an error signal to the satellite's attitude control and navigation system to help maintain a precise attitude and position estimate. In the ballistic missile application, the star tracker takes a fix on one or more stars, allowing errors in the launch position estimate to be determined and compensated for. The American Trident system and the Soviet SS-N-8 and SS-N-18 all used different 'stellar-inertial guidance' methods. The Soviets deployed this technology before the Americans [4].

Obstacle avoidance

Obstacles such as pylons and cables present a serious hazard to low-flying helicopters and aircraft. Obstacle databases can be used to generate warning symbols on head-up displays to indicate known permanent obstacles. Passive sensors will help the pilot to spot temporary and smaller features. Neither method will warn of thin temporary wires and cables.

Obstacle warning laser radar systems have been developed to fill this gap, and have been demonstrated for both helicopters and fixed wing aircraft. Helicopters fly relatively slowly ($30-100\,\mathrm{m\,s^{-1}}$) and fairly short-range systems will give sufficient warning for pilots to take appropriate avoiding action. Practical systems are available using diode lasers with obstacle detection ranges of a few 100 m. Fast jets need longer range warning and faster scanning. Demonstrator systems have been built using CO_2 lasers with coherent detection. These work, but are currently too large to integrate easily into fast jets.

C3.3.5.6 Training

Effective and realistic training is key to the effectiveness of modern military forces. If they cannot train in realistic combat conditions, they will not be effective when they first experience real combat. But if the training is too realistic, for example with a lot of live ammunition, not all the trainees would survive the experience!

So there is a need to make training as realistic as possible without using live ammunition. Traditionally, 'umpires' would move around training areas, making fairly arbitrary judgements about whether troops would have been hit. This was subjective and unrealistic. Similarly, blank ammunition produces noise and smoke, but the people doing the shooting get no feedback as to whether they pointed their weapons in the right direction. And if their equipment includes other hazardous systems, such as noneyesafe lasers which cannot be used in a training environment, they are unlikely to use them effectively under the stress of real war.

The need for effective training created a whole new class of EO systems, the 'direct fire weapon effect simulator'. A low power laser diode transmits a coded pulse train which if aimed correctly is detected by a detector on the target (or sometimes is returned to the transmitter by a retroreflector on the target and detected by a co-aligned receiver). The coding indicates the type of weapon being simulated, the divergence can be adjusted to simulate area effects, and round trip systems can measure distance to the target and therefore calculate whether the system was within effective range. Modern training systems inject video symbology into the sights of the attacking system to simulate weapon trajectories, tracer, and impact signatures; can electronically disable the weapons of the target if it has been 'killed'; and can keep a log of time and position of every engagement so that the engagement can be reconstructed in a virtual reality replay for debriefing and post-operation analysis. So laser training systems can be used to prove out the effectiveness (or otherwise) of new doctrine and tactics as well as to train soldiers. The systems contain detailed weapon effect and target information to ensure that engagements are correctly registered (e.g. a rifle cannot kill a tank).

These systems have immense entertainment value, and lightweight short range versions have become popular in the last few years in entertainment arcades!

Another class of training system is the crew training simulator, where very wide FOV video systems are used to generate a synthetic environment as seen from an aircraft cockpit or vehicle, which itself is on a 6-axis motion platform. The combination of physical motion and visual cues from the projected video creates a remarkably realistic experience. These systems save users massive amounts of money by avoiding the need to use real aircraft for many training tasks. Again, derivatives of these systems are now well established in the entertainment industry.

Training requirements reflect back on the specification of tactical military systems. For example laser rangefinders are often required to be ‘eyesafe’, so that they can be operated in accordance with operational drills in a training context; and similarly, modern fire control systems are required to be compatible with laser training systems and sometimes to incorporate various embedded training capabilities.

C3.3.5.7 Other applications

Covert communication

Optical methods have been used for line of sight communication since the discovery of the heliograph, probably in prehistoric times. Signal flags were used by navies in the days of sail. Once electricity was invented, signalling lamps allowed effective signalling at night with Morse code while maintaining radio silence.

During World War 2, there was a need for covert line of sight signalling, particularly for special operations. IR filters and image converters (the precursors of modern image intensifier tubes) were evolved for this purpose by the UK, Germany and USA. Nowadays similar principles can still be used. CO2 lasers can be used as beacons in the 8–12 μm region, near IR lasers are often fitted to weapons and surveillance turrets and used as ‘laser pointers’ to indicate targets to soldiers equipped with night vision goggles, and optical point-to-point data links have been developed which allow high bandwidth data to be sent over any clear line of sight—in the extreme case, between low earth orbit and geo-stationary satellites.

Infantry devices

Infantry soldiers use hand-held, weapon mounted and more recently helmet mounted EO sensors for observation, night vision, and weapon aiming. These have to be simple, compact, low power, and very robust to mis-handling.

A number of countries, including USA, France, UK, Canada, Australia and others, are now running ‘future soldier’ programmes which aim to integrate EO sensors, navigation, communications, command and control and protection functions into a ‘system’ suitable and affordable for use by infantry. The US Land Warrior programme, the British FIST, and the French ‘Felin’ have all demonstrated hardware and at the time of writing are moving towards full development and production.

All-round vision systems

All-round vision systems for military vehicles are an emerging market given that low cost sensitive sensors and flat panel displays are now available. The trend to smaller more compact platforms is creating a demand from designers to remove the vehicle design constraints (such as manned turrets in vehicles, bubble cockpit canopies in aircraft) imposed by the crew's expectation of good all round vision. The fear of possible future laser threats is creating a demand from operators for better situational awareness from indirect sensors and when vehicle hatches are closed down.

The DAIRS system for the Joint Strike Fighter (JSF) reportedly uses six 1000 × 1000 elements thermal imaging arrays to provide all round vision for the pilot and to provide data for threat warning and target detection systems. Similar proposals for land vehicles are less advanced but are likely to become reality in the foreseeable future.

Identification friend or foe

Identification friend or foe is a complex and difficult problem. Western military organizations in particular are increasingly expected to operate in coalitions with many different nationalities, with different types of equipment used on the same side—and sometimes the same kind of equipment used on opposing sides. Better visual identification makes a very important contribution to reducing 'friendly fire' incidents but cannot prevent them completely.

Many procedural and technical solutions have been tried over the years. In World War 2, distinctive black and white stripes were painted on all allied aircraft just before the D-Day invasion (of Normandy) to allow 'instant' recognition of a friend without need to positively identify the aircraft type. On a similar principle, active beacons matched to EO sensors are often used nowadays to give a distinct signature. These methods are very effective in specific operations, but do not provide a permanent solution, since over time they can be copied.

Laser interrogation is used as an operating principle in some modern combat identification systems.

C3.3.6 Operational impact

Thermal Imaging and night vision can 'turn night into day'—allowing forces to operate 24 h a day instead of 12, without having to use special tactics for night operation; or indeed to work preferentially at night to exploit the superiority resulting from thermal imaging and night vision technology. The US/allies' EO capabilities provided a huge advantage in the ground battle during the 1991 Gulf War.

Laser designation allows precision targets to be hit (fairly) reliably—about 30% missed in the Gulf War. There is no evidence that 100% accuracy can be achieved in operational systems; too many environmental and human factors influence the operational outcome. But precision weapons give a massive reduction in circular error probability (CEP), resulting in much lower mission costs and much less collateral damage to achieve a given objective. According to a quote in a web download:

> "On May 10, 1972 the first major combat use of these new weapons, along with TV guided Walleye bombs, took place (in Vietnam) resulting in the destruction of the Paul Doumer bridge, which had withstood 3 years of aerial bombardment, totalling 1250 tons of munitions. This single mission was a revolution in air-to-ground warfare."

The '1250 tons of munitions' had all landed somewhere near the bridge, devastating the surrounding area, and probably involving around 1000 operational sorties. We shall not attempt to pre-empt readers' own judgement about the environmental and human costs and moral issues. But it is clear that there is a very large financial payback from the use of precision weapons in terms of reduced cost of wasted ammunition, reduced cost of aviation fuel, and reduced combat losses. Similarly, laser rangefinders give a direct cost saving to the user calculable from the reduced number of rounds required to achieve a given number of hits, and the consequent savings throughout the logistics chain. Other EO technologies are less widely deployed—because the return on investment for the user is less clear and less easy to calculate.

These ‘value for money’ arguments will assure the place of EO in military inventories for the foreseeable future. Equally, interesting technologies which cannot provide similarly robust financial justification will have a much harder time gaining interest, let alone acceptance into service, from the military.

C3.3.7 Trends

Performance of EO systems is reaching operationally useful limits in many areas—so emphasis is switching from the ‘performance at any cost’ focus of the cold war years.

There is a strong drive to lower cost to allow wider deployment—for example thermal imaging is migrating from tanks (1985–1990 in the UK) to all combat vehicles (2000–2010 in the UK) to the individual infantryman (2000 onwards). This trend is aided by and is accelerating technology developments that allow simpler and cheaper systems to be developed, using staring and often uncooled IR detectors.

There is a strong drive to integration with other systems to reduce workload (Battle Management Systems, ‘soldier systems’); and to large-scale system integration throughout the ‘battlespace’ to improve ‘tempo’ or speed of operations, under the banner of ‘C4ISR’—command, communication, control, computing, information, surveillance and reconnaissance’. EO sensor systems (aided by their operators) will have to produce ‘information’ not just ‘images’.

The increasing use of unmanned vehicles will increase the demand for lower cost light weight electro-optic sensor systems, and for intelligent sensor integration and processing on the remote vehicles to reduce the demand for high datalink bandwidth between remote vehicles and operators by automating some of the operator’s traditional functions.

Some new areas of R&D now receiving much attention are:

(1) High power solid state lasers for directed energy.

(2) Mid IR lasers for IRCM.

(3) Three-dimensional imaging laser radar for target identification.

(4) Hyper spectral imaging.

(5) Mine detection using thermal imagers and active imaging.

And finally, commercial technologies are moving faster than military ones in the areas of telecoms, image sensors (perhaps except for 1–12 μm IR), and processing: the military are now exploiting commercial technology instead of the reverse. This may lead to the widespread use of ultra-low cost disposable EO systems on miniature autonomous vehicles and aircraft.

References

[1] Green R L 1958 *Tales of the Greek Heros* (London: Penguin)
[2] Wells H G 1898 *War of the Worlds*
[3] Jones R V 1978 *Most Secret War* (Hamish Hamilton)
[4] Mackenzie 1990 *Inventing Accuracy* (MIT)
[5] Chaisson E J 1994 *The Hubble Wars* (HarperCollins)
[6] Albedos http://www.drev.dnd.ca/poolpdf/e/61_e.pdf.2002
[7] Burns R W 1993 Early history of the proximity Fuze (1937–1940) *IEE Proc. A* **140**

[8] Flight International 2001 *Airborne Laser Designation*

[9] Reid B H 1983 The attack by illumination: the strange case of Canal Defence Lights *RUSI J*

[10] Sillitto H G 1999 Building a re-use infrastructure—optronics reference architecture and layered reference model *Proc. of INCOSE International symposium* (Brighton, 1999)

[11] Lloyd J M 1975 *Thermal Imaging Systems* (New York: Plenum)

[12] *Handbook of Artillery Instruments* 1914—HMSO 1914 (description of various optical rangefinders etc)

[13] Moss M and Russell I 1998 *Range and Vision (the first hundred years of Barr and Stroud)* (Edinburgh: Mainstream Publishing)

[14] Armstrong G, Oakley P J and Ranat B M 1993 Multi-mode IRST/FLIR design issues *SPIE Proceedings* CR **38**

[15] Netzer Y 1982 Line-of-sight steering and stabilization *Opt. Eng.* **21** (reviews various methods of line of sight stabilization and their suitability—fundamental techniques have not changed though improvements in gyro components and processing availability have changed some of the trade-offs!)

[16] Godfrey T E and Clark W M 1981 Boresighting of airborne laser designator systems *Opt. Eng.* **20** (A useful historical survey well describing many of the engineering issues)

[17] Jamieson T H 1984 Channel interaction in an optical fire control sight *Opt. Eng.* **23**

[18] *Handbook of Military Infra-red Technology* 2nd edn p 28

[19] Kopp C 1994 The sidewinder story—the evolution of the AIM-9 missile *Australian Aviation*

[20] Ward R D *Crossed Porro Resonator patent* ca 1974

[21] Gibson T *Night Vision Information* http://www.arctic-1.com/works2.htm

[22] McEwan K *Uncoooled IR Systems* various refs 1990–2002

[23] Pike J *Walleye guided bomb description* http://www.intellnet.org/documents/100/060/165.htm

[24] Rogers M E 1997 *USAF: Lasers in Space—Technological Options for Enhancing US Military Capabilities* Occasional Paper No. 2 (Center for Strategy and Technology, Air War College, Maxwell Air Force Base, Alabama)

[25] US Navy *Maverick Missile description* http://pma205.navair.navy.mil/pma2053 h/products/maverick/index.html

[26] Sofer B H and Lynch D K 1999 Some paradoxes, errors, and resolutions concerning the spectral optimization of human vision *Am. J. Phys.* **67**

[27] Bobbitt P *The Shield of Achilles* footnote p 204: reference to John Lynn, camouflage, *the reader's companion to military history*, p 68

[28] Sillitto R M 2002 Private conversations

[29] Vukobratovich D *Optomechanical Design Course Notes* SIRA, various years, CAPT, 2002

[30] Yoder P 1992 *Optomechanical Design*

[31] Mitchell C 1970 *Having Been a soldier* (London: Mayflower Paperbacks)

C3.4
Optical information storage and recovery

Susanna Orlic

C3.4.1 Introduction

Optical data storage has a long history dating back to the 1960s but, with the compact disk initially, it became relevant to the consumer and industry. The success of the early laser disk indicated the possibility of data storage based on optical phenomena and materials as an alternative to magnetic storage. Optical storage offers reliable and removable storage media with excellent robustness and archival lifetime and very low cost. Today, optical disk technology covers a wide variety of applications ranging from content distribution to professional storage applications. One of the major application areas for optical storage disks is the secondary storage of computer data in PCs and computer networks. An optical storage system is a particularly attractive component of the data storage network because it provides fast data access times and fair storage capacities while serving as a link between different multimedia and computerized systems. Perhaps, the most enabling feature of optical storage is the removability of the storage medium which allows transportation and exchange of the stored information between desktop and laptop computers, audio, video players and recorders. In contrast to the flying head of a hard disk drive, there are separations of a few millimetres between the recording surface and the optical ‘head’ while active servo systems enable dynamic recording and readout from a rotating disk. Consequently, the medium can be removed and replaced with relatively loose tolerances allowing an optical disk to be handled in different drives.

The Compact Disc digital audio has been launched into the era of digital entertainment by Sony and Philips in 1982. CD has enjoyed unprecedented success and universal support among electronic companies and hardware manufacturers. From its origin as the music storage medium for entertainment, compact disk has grown to encompass computer applications (CD-ROM), imaging applications (CD Photo) and video game applications (CD Video). The CD-ROM drives became standard in personal computers and CD-ROMs enhanced the efficiency of distribution and use of software, games and video. These read-only disks containing 680 MB of information can be mass replicated by injection moulding in a few seconds, and they are virtually indestructible.

Significant advances in the enabling technologies constituting the compact disk make it possible to increase capacity in digital-versatile-disk (DVD) format with the ability to store an entire movie in high-quality digital video on a single disk. With the DVD, optical disk storage opens a new chapter for video and multimedia applications. Laser optics, thin film and disk replication technologies have made considerable strides in the last decade. Digital coding and compression algorithms have become more sophisticated; integrated circuits and drive mechanisms have also advanced. Furthermore, the intensive research on new optical materials and recording phenomena has resulted in development of new technologies. The CD-Recordable, a recordable write-once system, and the CD-Rewritable drives and media have been introduced to the optical disk market. With its recordable formats, either for single recording or for expanded recording capabilities as in the case of CD-RW, optical storage became an alternative to the established hard disk and floppy disk systems, which dominated the data storage industry.

Versatility of optical storage is an additional feature of foremost importance for its applicability. Optical recording allows read-only, write-once and rewritable (erasable) data storage. Read-only optical disks or read-only-memory (ROMs) are suitable for distribution of digital contents. In read-only technology, the information is written on a master disk that is then used for printing the embossed patterns onto a plastic substrate. The printing process allows for rapid, low-cost mass reproduction making optical disks the media of choice for distribution of digital data. The write-once read-many (WORM) technology allows one to store permanently a large amount of data on a thin disk medium, to remove it, and to have fast access to it in any compatible optical drive system. Information stored on rewritable optical disks can be erased and rewritten many times. Present optical drives are designed to handle different media formats at the same time—the data stored on read-only, WORM and rewritable media can be accessed all in one unit.

Optical information storage is based on laser-material interaction for writing and reading. More general, data storage is the conversion of raw information into physical changes in an appropriate recording medium. Data recovery is the recognition of the stored information from the storage-induced changes. The physical processes applied to different storage technologies differ widely and define their performance limits and applicability. In conventional technologies, optical storage relies on data recording on a rotating, disk-shaped medium while recovery is done by optical effects. Typically, a focused light beam is used to interact with the physical structure of the data stored on the disk. The interaction between the read beam and the medium can be due to a number of different effects but it must modify some detectable light property. The light reflected from the medium is modulated by the stored data and is then directed to a photodetector that converts the optical signal into electronic signals. All conventional optical disk systems are based on the reflection mode but often use different recording and readout mechanisms. There is a variety of optical effects and materials which have been investigated and employed in various approaches to realize an optical memory. Table C3.4.1 summarizes different recording mechanisms and corresponding data structure for readout in well-established optical storage concepts.

Further advances in optical memories with high storage capacity up to terabyte require research on new photoactive materials exhibiting strong laser-induced changes of their optical properties. Such materials constitute a special class of nonlinear optical materials with optical properties depending on the incident light intensity or energy. This chapter deals with the fundamentals of optical storage but also with novel approaches to higher performance storage.

Table C3.4.1. Optical information storage: recording mechanism and corresponding data structure in different approaches.

Recording mechanism	Data storage/Readout method
Print	Surface grooves
Phase change	Reflectivity change
Magneto-optic	Polarization rotation
Photochromic	Complex refractive index change
Photorefractive	Refractive index grating

C3.4.2 Optical disk storage technologies

Removability of optical storage media is an attractive feature but made the standardizing efforts more complex compared with magnetic storage. Standardization in optical information storage was first accomplished with the appearance of CD creating the standards for Compact Disc Digital Audio format in 1982. Five years later, Sony and Philips announced the standard for compact disk storage of computer data—the CD-ROM. The establishment of worldwide standards has made it possible to incorporate writing technologies into CD systems. The recordable compact disks CD-R is functionally compatible with the standard CD-ROM, even though the reflectivity changes on a CD-R are induced by physical processes different from the interference effect at the printed grooves on a conventional compact disk.

Following to the CD-Recordable, CD-RW, the rewritable compact disk, has been introduced based on phase-change technology. Here, the reflective layers have two states which differ sufficiently in reflectivity to be read optically. Table C3.4.2 gives an overview of CD technologies starting with the CD-Audio, that was modified to allow new applications and later advanced to exploit the new technologies with CD-Recordable and CD-Rewritable.

In general, an optical disk storage systems consists of a drive unit and a rotating disk medium containing the information. The optical components are integrated within a small optical head allowing both optical recording and readout to be performed with the head positioned relatively far away from the storage medium, unlike magnetic hard drive heads. This allows the medium to be removable and effectively eliminates head crashes, increasing reliability. On the other hand, an optical head is heavier and leads to slower access times when compared to hard disk drives. The data recorded on an optical disk is organized in tracks that might be a single spiral or concentric rings spreading out along the radial direction.

In CD technologies, optical disks are either pre-recorded or pre-formatted with continuous grooves or a discontinuous groove structure that is needed to position the optical head on the disk. A laser beam focused to a small spot is used for recording and readout. Information is written bit-wise by modulating the properties of the recording material under illumination. During retrieval, recorded bits are detected as changes in some optical property of the light reflected from the disk. These changes might affect the amplitude, phase or polarization of light, and are sensed by a detector in the optical head. The rotation of the disk as well as positioning of the pick-up system is provided by a drive motor. In addition, the optical head contains a servo system based on some sort of optoelectronic feedback that is necessary to control the position of the focused laser spot on a rotating disk.

The optical storage systems introduced first worked as write-once, read-many systems. The information is recorded on a WORM disk by applying the laser beam to change the reflectivity. On the first ablation type WORM disks, the write laser beam burns a hole or pit into a thin absorbing layer, i.e. deposited aluminium. In the early WORM systems, different lasers have been used for writing

Table C3.4.2. Evolution of CD-based technologies from read-only memory to recordable and rewritable formats.

Disk Format	Recording Technology
CD Digital Audio	Read-only
CD Extra, CD Interactive CD-ROM, Video CD	Read-only
CD-Recordable	Write-once
CD-Rewritable	Rewritable

and reading. In a single laser system, the laser power is reduced for readout so that the recorded pits can be read by detecting the reflected energy. Following the ablation media, WORM disks with dye polymer media have been developed. Here, the laser beam induces a reflectivity change of dye polymer coating without material ablation. Rewritable storage has been first achieved in magneto-optical (MO) systems. The written information could be removed during an erase pass and thereby prepared for a new write cycle. To realize the MO systems, it was necessary to introduce new coating media and adapted optical systems. Consequently, there was no compatibility between WORM and MO storage media.

C3.4.2.1 Compact disk technologies

The basic concept of an optical disk involves a reflecting layer modulated by the presence of pits that then switch the reflected signal by destructive interference. In a CD-ROM drive the data is read out by projecting a focused laser spot on the rotating disk, then detecting the reflected light. The detected signal fluctuates with the presence or absence of pits along the track corresponding to the bit-wise stored information. Similarly, the recorded pits of CD-R and CD-RW are read out by detecting laser light reflected back from the disk. On the simple CD-ROM, the information is printed as quarterwave surface steps leading to zero reflectivity by interference. Tracks consist of discontinuous grooves or pits that are separated by lands whereby the pit length is determined by the bit content stored in it. The light reflected from the disk will therefore be modulated by the pits. Readout is done by destructive interference between the light reflected from the pit and the light reflected from the land. In this way, the data content represented by a pit sequence is translated into the reflectivity changes that occur when this sequence is exposed to the laser beam.

On a CD-ROM, pits are pre-formatted into the optical substrate (typically polycarbonate) that is then coated with a high-reflective alloy (e.g. aluminium or silver). The substrate itself entails the format that defines the physical structure of the data recorded on the disk. Present optical disk technologies use different recording mechanisms and are also based on different disk formats. In the simplest case such as CD or DVD-ROM, the disk have a static physical structure represented by pits. Other disk formats can be categorized in two types, one of which relies on continuous grooves, and the other on wobble marks. Different disk formats also require different fabrication processes for mass production. The well-known mass-production technique originates from the CD-ROM technology and is based on disk replication by stamping out copies of a disk master. This original disk is written using a short wavelength laser that creates the pattern in a photoreactive material. The typical mastering process is based on similar photolithographic techniques as applied in the semiconductor technology. Exposing the photoresist yields the pit structure, which allows for forming the so-called stamper. The final stamper comprises a metallic layer (usually nickel) with the inverted data pattern structure, which is then used for mass disk replication. In the production of ROM disks, stamping is followed by only two additional steps—coating with a reflective layer followed by a protective layer. In contrast, recordable and rewritable disks are more complex in structure as an optically active layer is needed for data recording. Depending on the technology, different materials and layer stack configurations have been developed to provide both the functionality and compatibility.

On the write-once CD-R, the focused laser beam locally and permanently changes the complex refractive index of an organic dye polymer layer (figure C3.4.1). Readout is based on signal enhancement or decrease by optical interference effects in the multilayer structure. The specification of the optical characteristics and thickness of the dye polymer layer allows the signal reflected from the reflective layer to be significantly decreased by the optical change induced in the polymer layer. The reflectivity difference between recorded 1's and 0's of CD-R and simple CD-ROM media is similar thus ensuring media interchangeability.

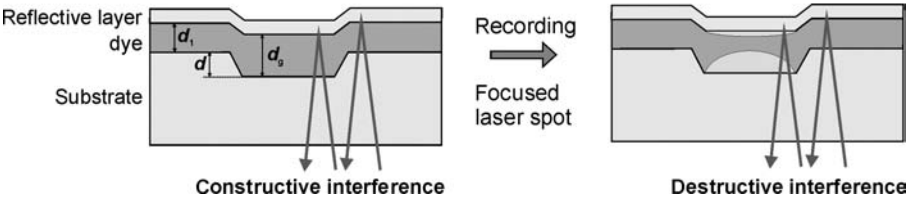


Figure C3.4.1. CD-Recordable: During recording, the pit is created by heating the dye absorptive layer with a focused laser beam.

Rewritable optical storage is presently based on MO or phase-change (PC) media. The compact disk rewritable drives and media introduced in 1997 by Philips represented a major breakthrough in optical disk storage. Using phase-change technology, the CD-RW systems allow disks to be written and rewritten many times over. CD-RW is therefore a medium of choice for both temporary and long-term data storage.

The disk structure in the case of phase-change media is more complex than the simple three-layer structure of a CD-ROM. The CD-RW disk consists of a grooved polycarbonate substrate onto which a stack of thin layers is sputtered, followed by a protective lacquer (figure C3.4.2). The phase-change or recording layer is sandwiched between two dielectric layers. These are typically zinc selenide–silicon dioxide ($ZnS{-}SiO_2$) layers which provide thermal tuning of the recording layer. The material for erasable phase-change recording is typically a $Ge_2Sb_2Te_5$ alloy which is sputter-deposited on a plastic substrate, with an undercoat and an overcoat of dielectric layers. Other phase-change materials such as $PdTeO_x$, InSnSb, AgInSbTe, etc, are also in use for both write-once and rewritable media. In addition, the stack comprises an aluminium layer from which the laser beam is reflected. As the disk comes out of the sputtering machine, the recording layer is in an amorphous state. The disk is then put into an initializer, which heats up the phase-change layer to the point where it crystallizes. Prior to recording, the phase-change layer is polycrystalline in its original state.

The phase-change recording mechanism relies on the reflectivity difference between dark amorphous zones and bright crystalline zones. For writing, a focused laser beam selectively heats small areas of the phase-change material near or above its melting temperature which is between 500 and 700°C. By heating the liquid state is achieved in the area under the laser beam spot. If the material is then cooled sufficiently quickly, the random liquid state is 'frozen-in' resulting in the amorphous state. To erase the recorded data marks, the laser heats the phase-change layer to re-crystallize the material. Hereby, the temperature of the phase-change layer is kept below the melting temperature but above

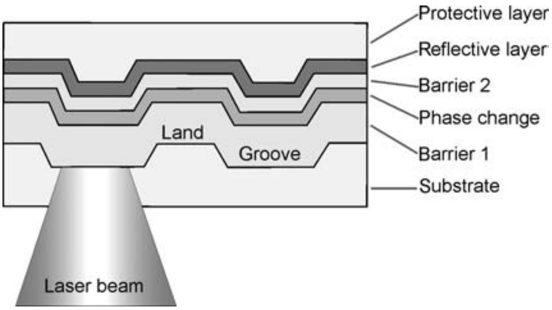


Figure C3.4.2. Structure of a CD-Rewritable: The active phase-change layer is surrounded by two dielectric barriers and covered by the reflective layer and protective overcoat.

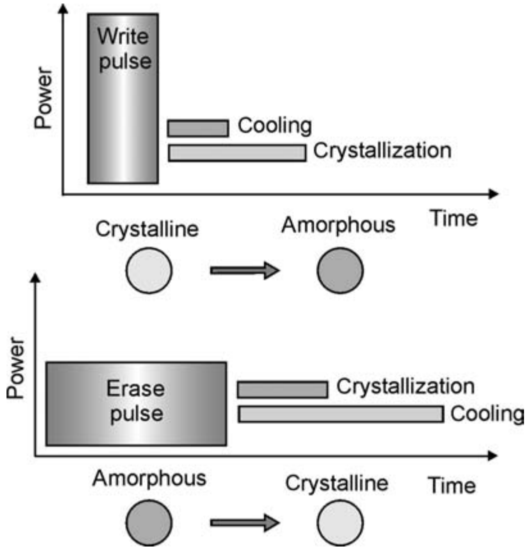


Figure C3.4.3. For writing, the phase-change layer is heated by a high-power laser beam to form an amorphous pit. The data is erased with a low-power laser beam by reverting the material to the crystalline phase.

the crystallization temperature for longer than the crystallization time so that the atoms revert back to an ordered, crystalline state (figure C3.4.3). The mark will become crystalline once again if the laser beam strikes the amorphous mark for sufficient time. Subsequent rewriting is possible when the material reverts back to an ordered crystalline state. The amorphous and crystalline states have different values of refractive index which tune the reflectivity of a multilayer stack. The recorded data can therefore be read out optically as reflectivity change between the low-reflective amorphous pits and high-reflective surrounding area. The phase-change process is reversible many thousands of times. Because the reflectivity difference of a CD-RW is much lower than the 70% of a CD-ROM, the sensitivity of CD drives has to be increased to allow readout of CD-RW media. In recent years, significant improvements have been achieved in both rewritable media and drive performance in order to increase the recording speed and cyclability.

C3.4.2.2 Digital versatile disk

Stimulated by the success of the CD, significant developments in many related fields have been undertaken resulting in further improvements. The large potential of accumulated R&D in that area made it possible to introduce the standard for the advanced DVD format. The basic structure of DVD-ROM is similar to the conventional CD-ROM but many parameters have been refined or reinvented to increase the surface data density and thereby the storage capacity.

The surface data density on an optical disk is inversely related to the spot size of the addressing beam. By reducing the spot size of the focused laser beam, smaller pits can be resolved and the surface density increases. This can be achieved by shortening the wavelength of the laser light and/or by increasing the numerical aperture of the optical system because the spot size is related to the wavelength divided by the numerical aperture. The specification employed to expand DVD's storage capacity includes smaller pit dimensions, more closely spaced tracks and shorter-wavelength lasers. Both the track pitch and the shortest pit length are nearly a half less than those of the CD (figure C3.4.4). The optical system has also been refined with a higher numerical aperture lens, resulting in a more tightly

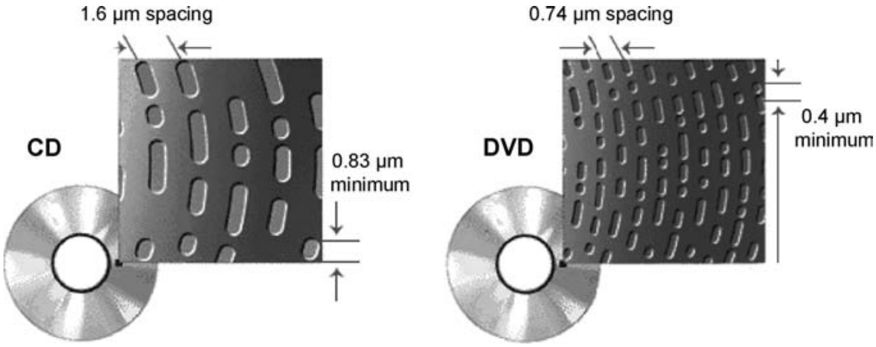


Figure C3.4.4. Pit-land structure of the CD compared to the refined data structure of the DVD (Sony).

focused laser beam. The simple single-layer DVD stores 4.7 GB per layer which is seven times the storage capacity of the CD.

Improvements in the overall storage capacity per disk have been achieved in different DVD format extensions that use two layers or both sides of the disk. However, due to the presence of the reflective layer, the number of layers per disk side is strongly limited to two layers; one of them is semi-reflective, and the other is high-reflective. The optical signal degradation when the reading laser beam passes through the first, semi-reflective layer as well as crosstalk effects between the layers require two distinct wavelengths, i.e. lasers, one for each layer. In general, dual-layer or double-sided configurations can provide moderate increases in storage capacity but, at the same time, they require more complex and more expensive systems.

Dual-layer, single-sided DVD is developed as a two-layer structure with the total storage capacity of 8.5 GB. The first semi-reflective layer reflects 18–30% of the laser light which is enough for the readout of the stored bits. With the transmitted light, the information from a highly reflective layer can also be read out. Gold was previously used as semi-reflective layer material, but new silicon layers as gold replacement provide significantly reduced production costs. Double-sided dual-layer DVD promises 17 GB but the electronics for reading and decoding the multilayer DVDs becomes much more complicated. An overview of different multilayer DVD formats is given in figure C3.4.5.

A similar two-layer structure is used in Super Audio Compact Disc, the new generation of digital audio media realized as a refined version of the CD audio to provide really high fidelity playback. This is done by adding one semi-reflective layer to the conventional CD structure that contains high density information. A silicon-based high density layer is semi-reflective at 650 nm wavelength and has almost 100% transmittance at 780 nm thereby allowing the standard CD laser light to be reflected from metallic reflective layer. With 650 nm wavelength, the pits smaller than those on a standard CD can be readout, leading to an enhanced storage capacity of up to 4.7 GB.

In recent years, implementation of recordable/rewritable technology in the DVD family has been the most serious challenge. Although, first recordable (DVD-R) and rewritable (DVD-RAM) products are available now, it is still not clear which format will become the standard DVD recording technology. Second generation recordable DVD format boosts capacities from 2.6 GB per side to 4.7 GB. However, rewritable DVD formats remain far away from achieving standards of compatibility, which have been the crucial issue for a successful introduction of CD-R and CD-RW. Three different DVD rewritable formats are presently competing for acceptance in the market place. The leading candidates to become the standard DVD recording technology—Sony's DVD + RW and Pioneer's DVD-RW—are incompatible so that the DVD family remains far away from CD harmony.

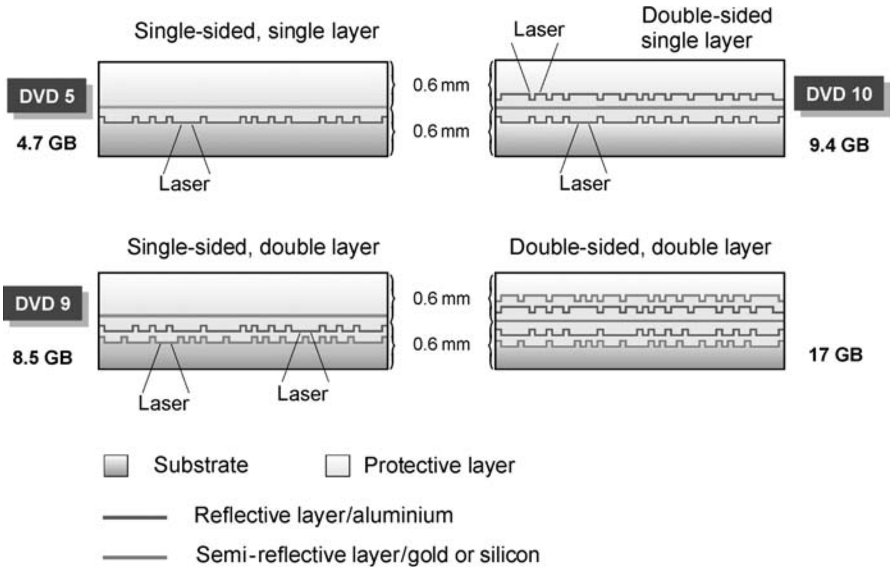

Single-sided, single layer
DVD 5
4.7 GB
Laser
0.6 mm
0.6 mm
Laser
Double-sided single layer
DVD 10
9.4 GB
Laser
Single-sided, double layer
DVD 9
8.5 GB
Laser
Laser
0.6 mm
0.6 mm
Double-sided, double layer
17 GB
Substrate
Protective layer
Reflective layer/aluminium
Semi-reflective layer/gold or silicon

Figure C3.4.5. Different DVD configurations and corresponding storage capacities.

The implementation of blue-violet diode lasers in DVD systems leads to a further increase of storage density. In February 2002 nine electronics companies (Philips, Sony, Pioneer and others) have established the basic specification for a forthcoming optical disk video recording format called ‘Blu-ray Disc’. Blue generation DVDs will store six times more data than current disks according to specifications announced by the consortium. Using a 405 nm semiconductor laser, the new video format enables 27 GB to be stored on a single disk. The consortium is also aiming to enhance the format by developing dual-layer disks with a capacity of up to 50 GB. The major application area is HDTV (High Definition TV) while computer storage is still under consideration.

C3.4.2.3 Magneto-optical disks

MO disks were the first rewritable storage media, available on the market since the beginning of the 1990s. MO disks are mainly used in professional data processing but with lower costs per megabyte, new applications are being opened. For consumer-oriented applications, the Mini Disc, a MO disk in the 2.5 in format was introduced as mini portable audio medium. In the last years, the MO disk technology has been advanced continuously further to allow improved storage capacity and read/write performance.

Data storage in MO disks is based on opto-thermic magnetic effects. Information is stored as a magnetized state of the magneto-optic layer and will be read out as polarization change of the laser beam using the Kerr effect. The MO layer that stored the information as corresponding magnetized state is protected from oxidation by two dielectric barrier layers (SiN) (figure C3.4.6). Together with the reflective layer the barriers ensure an optical signal enhancement. Presently all commercially available MO disks are based on an amorphous terbium–iron–cobalt magnetic alloy. This material belongs to a class of materials known as the rare earth–transition metal alloys.

For recording of binary data, the MO layer is heated by the laser spot above the Curie temperature where its magnetic orientation is dissipated. When this spot cools, the new magnetic orientation is set by

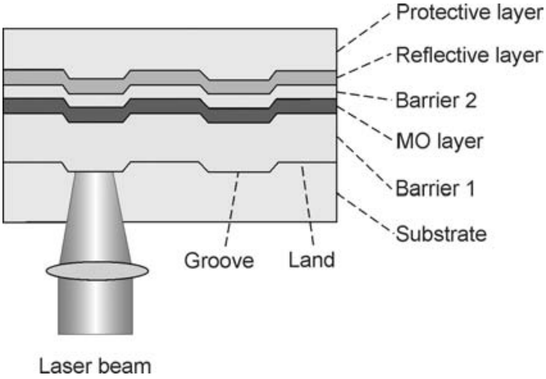

Protective layer
Reflective layer
Barrier 2
MO layer
Barrier 1
Substrate
Groove
Land
Laser beam

Figure C3.4.6. Structure of a magneto-optical disk.

the magnetic head corresponding to the '0' and '1' of the digital signal (figure C3.4.7). During readout, the polarization of the read beam is rotated, thereby detecting the recorded bits.

The MO media rotate the polarization vector of the incident beam upon reflection. This is known as the polar magneto-optical Kerr effect. The sense of polarization rotation is dependent on the state of magnetization of the medium. Thus, when the magnetization is pointing up, for example, the polarization rotation is clockwise, whereas down-magnetized domains rotate the polarization counterclockwise. The polar Kerr effect provides the mechanism for readout in MO disk data storage. Typical materials used today impart about 0.5° polarization rotation to the linearly polarized incident light. But, given the extremely low levels of noise in these media, the small Kerr signal nonetheless provides a sufficient signal-to-noise ratio for reliable readout. The media of MO recording are amorphous. Lack of crystallinity in these media makes their reflectivity extremely uniform, thereby reducing the fluctuations of the read signal as well as level of noise in readout.

To rewrite a MO disk, all previously recorded bits must be erased before new data can be recorded. This requires either a recorder with two lasers (one to erase and one to record) or longer recording times in a single laser system because the laser must erase the data in the first rotation and then record the new data in the second one. Alternatively, one can start with an erased track, apply a reverse-magnetizing DC magnetic field to the region of the interest, and modulate the laser power to record the information along the track. This is known as the laser power modulation recording

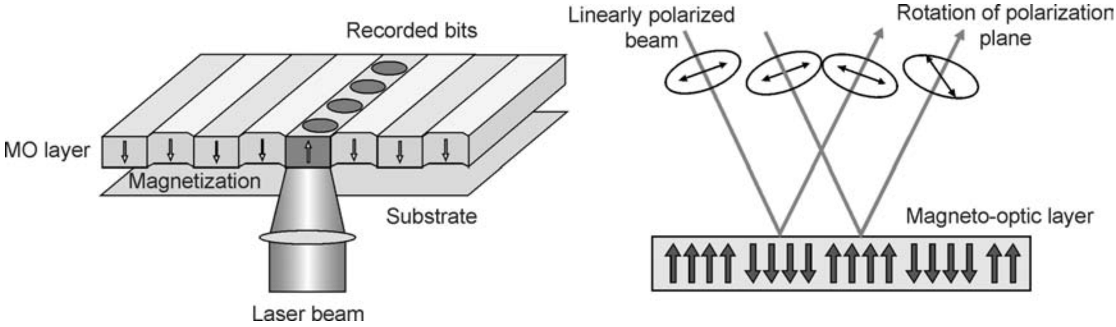

Recorded bits
MO layer
Magnetization
Substrate
Laser beam
Linearly polarized beam
Rotation of polarization plane
Magneto-optic layer

Figure C3.4.7. Left: Recording of a magneto-optical disk. The pit mark is formed by heating the MO layer with a focused laser light. Right: Readout of a magneto-optical disk. The data is read out by the polarization change of the read beam.

(or light intensity modulation—LIM) scheme. The laser power modulation does not allow direct overwrite of the pre-existing data on the track unless a more complex media structure is employed. Exchange-coupled magnetic multilayer structure allows LIM direct overwrite in magneto-optical systems. In such systems, the top and bottom layers are exchange coupled together so that switching one layer would make the other layer to switch, too. The top and bottom layer have perpendicular magnetic anisotropy and are separated by an intermediate layer which is in-plane magnetized and aimed to ease the transition between the top and the bottom layer. In an LIM direct overwrite scheme, the recorded domains collapse under a laser beam spot of moderate power. A high-power beam is used to create the domains through the whole thickness of the multilayer stack. The recorded domains are read out by a low-power beam, which do not disturb them so that only a laser beam of moderate power can erase pre-recorded marks. Erasure is very similar to writing, in that it uses the heat generated from the laser beam and requires assistance from an externally applied magnetic field to decide the direction of magnetization after cooling down. This method is simplified in a Mini Disc recorder based on a magnetic field modulation overwrite system. Here, the new data are written immediately over the previous data. Writing is achieved by a continuous-wave laser beam and a modulated magnetic field.

MO disks are currently available in 3.52 in single-sided and 5.25 in double-sided formats with storage capacities up to 640 MB or 2.6 GB, respectively. Storage densities of MO media are being substantially increased using new advanced technologies like magnetic super resolution (MSR) and near field recording (NFR). Doubling of storage capacity is expected for the first MO disks based on magnetic super resolution. Further advances resulting from MSR allow storage capacities of up to 10 GB for 5.25 in disks.

On MO disks, information is written by heating the recording material with the write laser beam. The stored data are retrieved by reflecting the read laser beam from the structured disk whereby the optical resolution is limited by diffraction. By using the MSR process, it is possible to write bits smaller than the laser spot, i.e. below the optical diffraction limits. This method is based on specific magnetic properties of the ‘rare earth/transition metal’ materials, which evolved as storage material for MO recording. The write process is the same as writing on conventional MO disks but the recorded bits are smaller than the laser spot. This becomes possible by using the temperature profile induced by the laser spot in a recording material (figure C3.4.8). The bits are written only within the central area of the laser spot where the magnetic material is reversed by its high temperature. The advantage of MSR is that the minimum mark length readable in an MSR system is shorter than any that are readable in conventional systems.

Comparing phase-change and MO technologies, one can find several advantages and disadvantages for each. For a number of reasons, phase-change recording appears to be the ideal solution for rewritable optical storage of information. CD-RW drives are simpler than MO drives, because they do not need magnets to create external magnetic fields, and because there is no need for sensitive polarization-detecting optics in CD-RW readout. Phase-change readout makes a profit from the reflectivity difference between crystalline and amorphous states which is large enough to provide much higher signals than the relatively weak MO effect. On the other hand, repeated melting, crystallization, and amorphization of phase-change media results in material segregation, formation of submicron areas that remain crystalline, stress buildup, etc. These factors might reduce data reliability and cyclability of the phase-change media. MO disks are guaranteed to sustain over 10^6 read/write/erase cycles while the corresponding figure for phase-change media is typically 1–2 orders of magnitude lower. The maximum temperature reached in MO media during recording and erasure is typically around 300°C, as compared to 600°C in PC media. The lower temperatures and the fact that magnetization reversal does not produce material fatigue provide the longer life and better cyclability of the MO media.

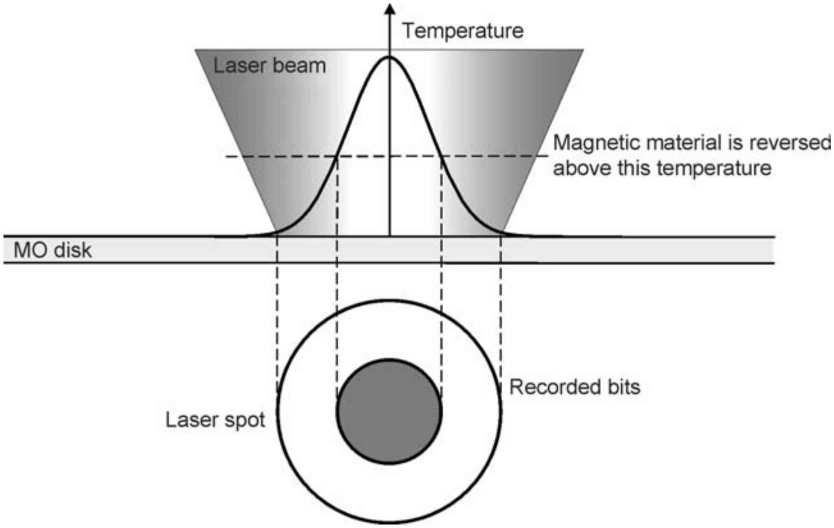


Figure C3.4.8. Magnetic super resolution. The temperature profile of a focused laser beam is used to create smaller pits than the laser spot.

C3.4.3 Optical disk storage system

In conventional, well-established technologies, an optical storage system consists of a drive unit and a storage medium, which is usually a rotating disk. The disk storage medium and the laser pickup head are rotated and positioned through a drive motor. In general, optical disks are pre-formatted by grooves and lands that define the so-called tracks on a disk. A track represents an area along which the information is stored; an optical disk consists of a number of tracks that may be concentric rings of a certain track width separated by a land area. The separation between two neighbouring tracks is the so-called track-pitch. Typical track spacing on existing optical disks are on the order of micron. In the simplest recording scheme, pit marks of equal length are created along the track (either recorded or stamped) while presence or absence of these marks corresponds to binary digits 0 and 1. Tracks might be physically existent prior to the recording or created by prerecorded data marks themselves.

On a read-only medium, for example CD audio disk, quarterwave deep pits are stamped into an optical disk substrate and then coated with a reflective layer to provide the readout by interference effects in the reflection configuration. In this case, pit marks define their tracks that are discontinuous grooves consisting of pits separated by lands. The length of both pits and lands within a track is defined by the encoded bit stream stored within that area. Tracks are necessary to enable the positioning of the optical head which accesses the information on the disk and to guide the laser beam during readout. In case of recordable media, the disk is typically pregrooved, i.e. continuous grooves are printed, etched or moulded onto the substrate to define pre-existing tracks. The grooves represent tracks and are separated by lands. The information is recorded along the tracks that can be either concentric rings or a single spiral. Alternatively, the lands may also be used for recording—in this case, adjacent tracks are separated by grooves. Moreover, land-groove recording, i.e. recording in both land and groove areas, has been introduced in DVD technology to increase storage density by a more efficient usage of the available storage area.

All optical components needed for recording and retrieval of information are integrated in an optical pick-up system or simply optical head. Today's optical heads are small, compact and highly optimized systems that fly close to the disk surface. The optical head must be able to rapidly access any position on the disk for error-free recovery. Depending on the disk technology, the optical head might have different architectures but the basic configuration is the same for all present systems. Usually, the head comprises a laser diode, a collimator lens, an objective lens, a polarizing beam splitter, a quarterwave plate, and the detector system. A typical setup for an optical disk system is shown in figure C3.4.9 [1]. A linearly polarized laser beam is emitted by the laser diode. The collimator reduces the divergence and collimates the beam, which then passes through a polarizing beam splitter and a quarterwave plate. The quarterwave plate circularly polarizes the incident beam, which is then focused onto the disk by an objective lens that also collects the light reflected back from the disk. For recording, a focused laser beam generates a small spot within the active material to induce some kind of optically detectable changes. The recorded bit marks will then change the phase, amplitude or polarization of the readout beam. The total reflected light is therefore modulated by the presence of pits—the light fractions reflected from the pits and the light fractions reflected from the land interfere destructively and modulate the data signal according to the stored information. The beam reflected back from the disk again passes through the objective lens and becomes re-collimated afterwards. The quarterwave plate converts the circular polarization of the reflected beam to a linear one with the direction perpendicular to that of the incident beam. In this way, the incident, readout beam and the reflected, signal beam can be separated by the polarizing beam splitter which directs the reflected light to the data detector. The detection system produces the data or readout signal but also optoelectronic signals needed for automatic focusing and track following. Specific servo systems are required to control the position of the optical head with respect to the tracks on the disk. Additional peripheral electronic units are used for functional drive control, data reconstruction and encoding/decoding.

An optical disk storage system is characterized by several functional quantities that specify its performance in terms of capacity and speed. Typical parameters are storage capacity, access time, data transfer rate and cost. The storage capacity is determined by the areal density of the stored information and the geometrical dimensions of the disk medium, i.e. available storage area. The areal density characterizes the efficiency of a system in using the storage area and is a direct function of the spot size and/or the minimum dimension of a stored bit mark. It is typically given in units of gigabits per square inch or bits per square micron. The areal density is limited by the optical resolution of the laser pick-up, i.e. by the minimum dimensions of data marks that still can be detected by the optical system. The numerical aperture of the objective lens and the wavelength of the laser used for recording and readout

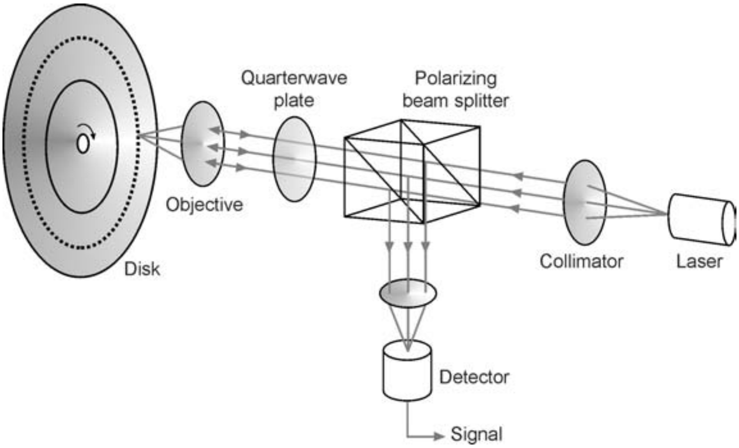


Figure C3.4.9. Basic configuration of an optical disk system.

determine the diffraction-limited spot size and therefore the data density but further factors also have to be taken into account such as track density and linear bit density. The track density (tracks/in) expresses how close to each other neighbouring tracks on the disk can be arranged. The linear bit density (bits/in) is a metric for the spacing between optical transitions along a track.

Other parameters of interest are access time and data transfer rate, which characterize how fast the information stored on the disk can be accessed and read out. The access time depends primarily on the speed with which the optical head can move over tracks to access a given storage location on the disk. The data rate is a metric for recording and readout speed and it depends on the linear bit density and the rotational speed of the drive. The data rate becomes important when large data files, for example images or video files, have to be stored or retrieved. Different paths to higher data rates have been proposed in CD/DVD technology including, for example, recording/readout by multiple beams, i.e. usage of several laser pickup systems simultaneously. The laser beam power available to write and read information, and the speed of the servo system controlling the optical head position are some of the parameters that limit the data transfer rate.

C3.4.3.1 The objective

The objective lens in an optical disk system must be designed to correct for spherical aberration which is due to the substrate thickness. Current optical disk systems use moulded glass lenses to focus the laser beam to a diffraction-limited spot. There are several advantages of the moulded lens which is typically a single aspheric lens over conventional objective lenses made up of multiple elements. These advantages result from the moulding process itself which is much better suited to mass fabrication. Furthermore, the working distance, i.e. air spacing between the objective lens and the disk surface, of a moulded lens is larger which simplifies the design and function of the head in a removable-media optical system. Also, the moulded lens mass is lower than a conventional lens which reduces actuator forces needed for automatic focusing and track following within the servo system.

In order to achieve ultimately small spots on the reflecting data layer, the objective lens must have fairly large numerical aperture, and it must be free from aberrations. Hereby, the numerical aperture (NA) of a lens is defined by $\mathrm{NA} = n \times \sin\theta$, where θ is the half angle closed by the cone of the focused laser beam. The diameter of the diffraction-limited focused spot is then given by the ratio of the wavelength of the laser beam, and the numerical aperture of the objective lens, i.e. $d_{\mathrm{spot}} \approx \lambda_0/\mathrm{NA}$, where λ_0 is the vacuum wavelength of the laser beam. It becomes clear from the above relation that higher numerical apertures are desirable if smaller spots and therefore higher storage densities are to be achieved. The smaller the spot of the readout beam, the smaller data marks can be resolved. The areal data storage density depends directly on the spot size and according to the above equation, it can be increased by reducing the wavelength and/or by increasing the numerical aperture. In practice, both approaches encounter limitations and also require adaptations and advances in a number of constituting technologies. The most important aspect in achieving greater storage densities by reducing the wavelength is the availability of short wavelength lasers.

Reducing the spot size by increasing the numerical aperture will also reduce the focal range of the laser beam, the so-called depth of focus. The focal range can be estimated by the Rayleigh length of the laser beam and is therefore proportional to λ/NA^2, which means that the higher NA, the smaller will be the depth of focus. That clearly limits NA as an optical storage system is capable of handling the focus error only within this range. Other tolerances such as those for the disk tilt and the substrate thickness are also limited by NA and have to be considered in designing the objective lens. Presently, the NA of objectives used in CD technology is 0.45, which has been increased to 0.6 in DVD systems to achieve higher storage capacity. A further increase of NA would require far-reaching adaptations in both the disk configuration and optical system to provide that the diffraction-limited focus will be maintained with the desired accuracy on the reflective data layer.

C3.4.3.2 The laser

Optical storage of information became first practically possible with the invention of laser. Rapid developments in the field of laser systems have supported the technological realization of existing optical storage systems. In general, lasers are used in all optical technologies for data recording and often also for data recovery. The ultimate premises for light sources in optical storage systems are small size, stability, long lifetime and inexpensive mass production. The optical drive so as we know it became therefore possible first with the establishment of laser diodes. Conventional optical disk systems, such as CD, DVD and MO, rely all on semiconductor laser diodes as their source of light. The diode laser technology has been the key enabling technology for optical storage while the success of CD consumer products (CD audio and CD-ROM in PCs) has pushed the diode laser to the best selling laser products of all time. Compared to other laser types, the laser diode has many advantages with regard to the requirements defined by the design and functionality of an optical head. In general, the optical head contains one laser diode that provides light for recording, reading and erasing (given in rewritable systems only). Each of these functions sets specific requirements on the light source.

Optical data storage systems for mainstream applications such as computer and entertainment need compact and cheap lasers that can be with ease integrated in low-weight optical heads and small, low-cost drives. In addition, specific requirements of optical data storage concern the wavelength of emitted light, optical power available for writing, reading and erasing, modulation, and life time of the laser diode which should exceed several thousand hours. The wavelength of light used to write and read bit marks is crucial for the optical resolution of the pick-up system and, consequently, for the areal storage density. The shorter the wavelength the smaller bits can be resolved and the higher areal storage density can be achieved. Stimulated by the success of CD and later DVD, extensive research and industrial development efforts have been undertaken in past decades to satisfy demand for short-wavelength laser diodes with sufficient optical power for the operation in optical drives.

The power requirement for diode lasers in optical disk systems varies from several milliwatts for retrieval of stored information to several tens of milliwatts for data recording. The laser output radiation is modulated directly by modulating the input electrical current. The fast modulation of laser radiation is perhaps the most important characteristic of laser diodes for optical recording. Laser diodes can be modulated to GHz frequencies with rise and fall times of less than 1 ns. For read-only applications, low-power lasers with approximately 5–10 mW optical power can be used. In contrast, for recordable and rewritable media more power is needed because the laser beam has to induce almost instantaneously detectable changes in the recording material. For example, amorphous pit marks on a CD-RW are thermally induced by a pulsed laser beam with optical power of either 50 or 60 mW.

Another requirement concerns the spatial coherence and single transverse mode operation as the laser beam has to be focused to the diffraction limit. The laser cavity must be a single mode waveguide over its operating power range to provide wavefronts that are needed to achieve ultimately required small spots. The longitudinal mode stability has not been a requirement up to now so that laser diodes incorporated in the present optical heads typically operate in several longitudinal modes. Otherwise, a single longitudinal mode operation might become important in reducing undesired wavelength fluctuations. Fluctuation of both intensity and wavelength is one of the characteristics of diode lasers. While intensity fluctuations reduce the signal-to-noise ratio of the readout process and generate noise in the servo signals, wavelength fluctuations set additional requirements on achromatic design of optical components.

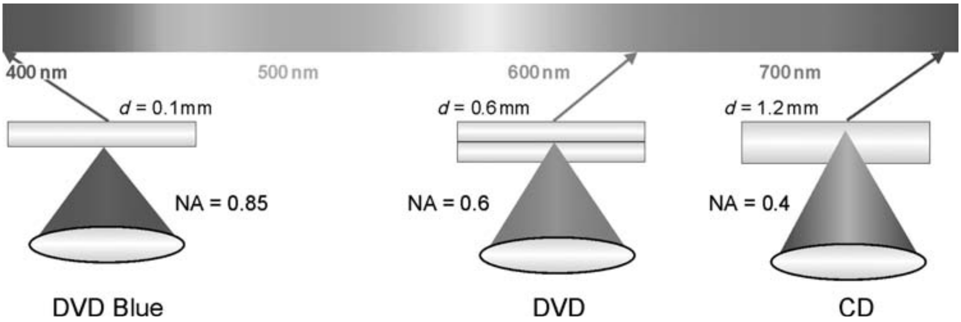


Figure C3.4.10. The increase in storage density from CD to DVD has been primarily achieved not only by decreasing the bit dimension but also by reducing the tracks spacing, by using shorter-wavelength lasers and higher numerical aperture optics. At the same time the substrate thickness has been reduced from 1.2 mm on a CD to 0.6 mm on a DVD, and new blue-generation DVDs will have a protective layer of 0.1 mm thickness only.

The wavelengths of optical data storage have been continuously reduced since the introduction of the first laser disk systems which started at 830 nm. CD audio and early CD-ROM systems rely on infrared diodes at 780 nm while moderate power lasers at 680 nm are used in CD-Recordable and MO products, and also in computer drives. The DVD-ROM standard relies on red-emitting diodes at 635 and 650 nm. The wavelength reduction has been one of crucial factors for the sevenfold increase in storage capacity from CD to DVD (figure C3.4.10). A shorter wavelength laser would support a higher storage density as shorter-wavelength light can be focused to a smaller spot at the diffraction limit. As the diameter of a focused laser spot is proportional to its wavelength, the reduction of the wavelength will lead to the reduction of the spot size by the same factor, and consequently, to an increase of the storage density by the square of that factor.

Recent developments in the field of III–V semiconductor diode lasers allow for an almost revolutionary transition from the red to the blue-violet spectral range. Enormous advances can be achieved in coming years by incorporating short-wavelength blue-violet diode lasers in optical storage systems. Blue-violet diode lasers operating in the range of 370–430 nm have been developed by Nichia Chemicals Corporation of Japan and introduced to the market in 1999. A new generation of DVD using blue laser diodes is expected to penetrate the market within the next 5 years. The standard recently announced by major optoelectronic companies involved in optical disk technologies is based on a 50 mW blue-violet diode laser. By adopting a 405 nm laser, the new DVD will minimize its beam spot size by increasing the numerical aperture to 0.85. On the other hand, such an extremely sharply focused laser beam is characterized by a very small focus depth. Therefore, the substrate becomes also extremely thin with a thickness of 0.1 mm only so that the disk must be protected from outer influences and damages by a cartridge. The both issues, short wavelength and high NA, allow for reducing the pit size to approximately 0.2 μm, and the tracking pitch to 0.32 μm, almost half of that of a regular red DVD. All these refinements and improvements together should push the DVD technology up to 27 GB high-density recording.

C3.4.3.3 The servo system

Readout in optical storage relies on data reconstruction from a disk rotating with a several thousand rotations per minute. To provide a faithful retrieval of the stored data, the laser beam must be focused exactly on the disk track and then maintained accurately within it during the entire readout process. Typical tolerances in optical disk systems are 1 μm for positioning of the focus in the reflective data layer, and one-tenth of micron for focus positioning on the track centre. The axial and radial runout of an optical disk are two or three order of magnitudes larger than these allowable focus positioning errors. The optical drive therefore requires a servo system to compensate the radial and vertical runout of the disk as it spins, and to provide submicron focus and track-locking schemes. The servo system is a closed-loop opto-electro-mechanical system which couples optical position detectors to high bandwidth actuators to actively follow the disk rotation. Consequently, the servo control in an optical disk system involves accurate and continuous focus position error sensing and sophisticated feedback mechanisms that dynamically convert the detected error signals in corresponding actuator movements.

The task of actuators within the servo system is to correct and to control the position of the focusing optics, i.e. objective lens. A typical objective lens has a numerical aperture of 0.45 or higher, to create a focused beam spot smaller than 1 μm. In addition, the focused beam has a focus depth which is only a fraction of a micron. On the other hand, a rapidly rotating disk has a tendency to wobble in and out of its ideal position in the optical drive. There is a variety of reasons that may cause such wobble effects, some of them are imperfections in the disk construction, substrates that are not ideally flat, other manufacturing errors in both disks and drives, disk tilt and eccentricity, etc. In the ideal case, the disk mounted in a drive would be perfectly centred, and sufficiently flat to maintain an ideal perpendicular position with respect to the rotation axis at all times. Deviations up to $\pm 100\,\mu\text{m}$ in both vertical and radial directions usually occur during the disk rotation in an optical drive.

The task of the servo system is twofold: First of all, the laser beam spot focused into the reflective layer of the disk must remain within the depth of focus, and second, the focused spot must remain within a submicron-sized track while the disk rotates and wobbles in and out of both focus and track. The mechanism to maintain the laser spot on the disk within the focus depth is called automatic focusing and it is aimed to compensate vertical runout of the disk. In addition, automatic tracking servo is needed to maintain the position of the focused spot within a particular track—the mechanism is called automatic track following. In the ideal case, tracks are perfectly circular and concentric, and the disk is perfectly centred on the rotation axis. In practice, the eccentricity of both the tracks and disk creates demand for active track following mechanisms that compensate wobble effects and enable the laser beam spot to follow the track.

Automatic focusing

Current optical drives rely on several different methods that provide the feedback mechanisms, i.e. error signals that drive the focus servo system. The objective lens is mounted in a voice coil actuator with a bandwidth of several kilohertz, and the feedback mechanism is used to position the lens relative to the rotating disk in such a way as to maintain focus at all time. The basic premise is that an appropriate error signal is generated which is then fed back to the voice coil actuator for maintaining focus automatically. Depending on the detection scheme, various techniques have been proposed to generate the focus error signal (FES). The signal needed for the closed servo loop is usually derived from the light that is reflected from the disk. Several techniques use a field lens that creates a secondary focused spot; deviations from optimum focus are then analysed by observing that secondary spot. The field lens which might be a spherical, ring-toric or astigmatic lens is placed after the objective lens in order to focus the light reflected back from the disk. The shape, size and position of the focused spot depend on the position of the disk relative to the in-focus plane. Changes in the secondary spot caused by the off-focus status of the disk are detected via a photodetector and transformed into an electronic signal that contains the feedback information for the focus servo system.

A very popular mechanism for automatic focusing relies on the astigmatic lens detection method. The most of current optical disk systems use an astigmatic servo sensor that comprises an astigmatic lens having two different focal lengths along orthogonal axes, and a quad detector. Figure C3.4.11 shows

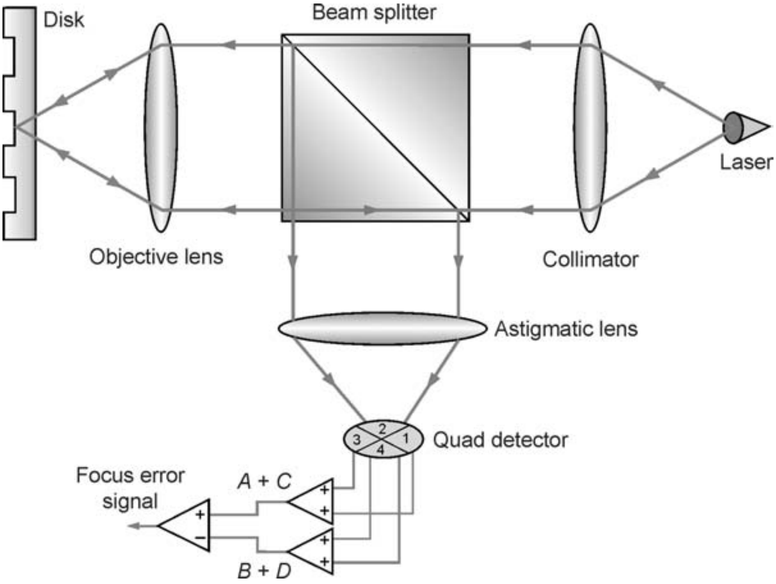

Disk
Beam splitter
Laser
Objective lens
Collimator
Astigmatic lens
2
3
1
4
Quad detector
Focus error signal
A + C
B + D
+
+
+
+
+
−

Figure C3.4.11. Astigmatic focus-error detection system.

a diagram of the astigmatic focus-error detection system used in many practical devices. The light beam returned from the disk and collimated by the objective lens might be convergent or divergent, dependent on position of the disk relative to the plane of best focus. The reflected beam passes an astigmatic lens, which normally focuses the incident light beam to a circularly symmetric spot halfway between its focal planes. In the best, in-focus case, a quad detector placed at this plane receives equal amounts of light on its four quadrants. In contrast, when the disk is out of focus, the astigmatic lens will create an elongated spot on the detector so that individual quadrants will be illuminated differently and, consequently, they will create different electronic signals (figure C3.4.12). Depending on the sign of defocus, this elongated

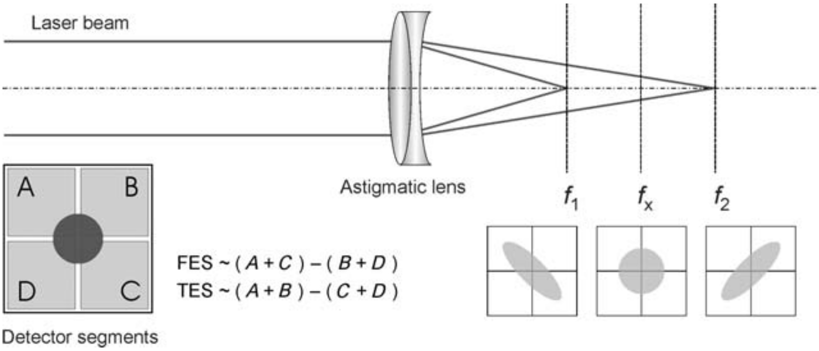

Laser beam
Astigmatic lens
f_1
f_x
f_2
A
B
D
C
FES ~ (A + C) − (B + D)
TES ~ (A + B) − (C + D)
Detector segments

Figure C3.4.12. Four-quadrant detector combined with an astigmatic lens is used in optical disk systems for both automatic focusing and track following. Control electronic signals are generated by an appropriate combination of the signals from four detector segments.

spot may preferentially illuminate quadrants A and C or quadrants B and D of the detector. A bipolar focus error signal can be then derived as the difference between diagonal quadrants, i.e.

$$\mathrm{FES} \propto \frac{(A+C)-(B+D)}{A+B+C+D}.$$

Automatic track following

The information is recorded on an optical disk either around a series of concentric circular tracks or on a continuous spiral. Manufacturing errors and disk eccentricities caused by mounting errors or thermal expansion of the substrate, for example, will cause a given track to wobble in and out of position as the disk spins. Typically, a given track might be as much as ±100 μm away from its intended position at any given time. The focused spot is only about 1 μm across and cannot be at the right place at all times. An automatic tracking scheme is therefore desired. The feedback signal for controlling the position of the objective lens within the tracking coil is again provided by the return beam itself. The four segments of the detector are combined in different ways for focus-error signal (FES) and track-error signal (TES). Several mechanisms for automatic track following have been proposed and applied in commercial devices.

The push–pull tracking mechanism relies on the presence of either grooves or a trackful of data on the media. In the case of CD and CD-ROM, the data is prestamped along a spiral on the substrate, and the sequence of data marks along the spiral represents a sort of discontinuous groove structure. Writable media such as CD-R, MO and PC require a tracking mechanism distinct from the data pattern, because prior to data recording, the write head must be able to follow the track before it can record anything. Once the data is recorded, the system will have a choice to follow either the original tracking mechanism or the recorded data pattern. Continuous grooves are the usual form of pre-existing tracks on optical media. A typical groove is a fraction of a micron wide and one-eighth of the wavelength deep. As long as the focused beam is centred on a track, diffraction of light from the adjacent grooves will be symmetric. The symmetry of the reflected beam, as sensed by a quad detector in the return path, would produce a zero error signal (figure C3.4.13). However, when the focused spot moves away from the centre of

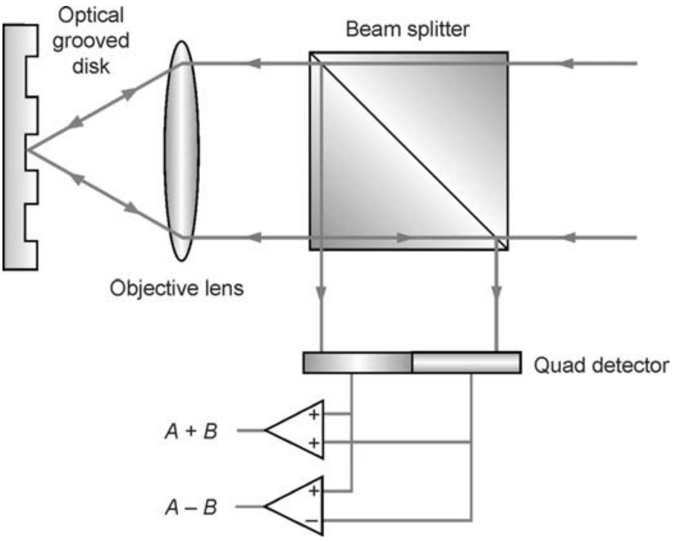


Figure C3.4.13. Track-error signal generated by push–pull method.

the track, an asymmetry appears in the intensity pattern at the detector. The difference signal is sufficient to return the focused spot to the center of the track. The track error signal is thus generated by the quad detector according to

$$\mathrm{TES} \propto \frac{(A+B)-(C+D)}{A+B+C+D}.$$

In read-only media, where continuous grooves are not present, another method of tracking is applied, the so-called three-beam method. The laser beam is divided into three beams, one of which represents the main spot and follows the track under consideration, while the other two are focused on adjacent tracks, immediately before and after the addressed track. Consequently, three detectors are necessary for position error sensing. Any movement of the central beam away from its desired position will cause an increase in the signal from one of the outriggers and, simultaneously, a decrease in the signal from the other outrigger. A comparison of the two outrigger signals provides the information for the track-following servo.

In another possible tracking scheme, the so-called sampled servo scheme, a set of discrete pairs of marks is placed on the media at regular intervals. The wobble marks indicate the transversal boundaries of a track. Such marks might be embossed or written by a laser beam within the formatting procedure. These marks are slightly offset from the track centre in opposite directions, so that the reflected light first indicates the arrival of one and then of the other wobble mark. The track error signal is generated as the difference between the signals detected from each wobble mark. Depending on the spot position on the track, one of these two pulses of reflected light may be stronger than the other, thus indicating the direction of track error. The sample servo technique is often used in recordable media systems including both write-once and rewritable media.

In contrast, present DVD systems rely on the differential phase detection technique that is based on diffraction of the focused spot from the edges of data marks. If the spot is focused off-track, the light reflected from the disk will show an asymmetric intensity distribution each time the spot strikes a mark edge. The intensity pattern rotates when the spot travels along the data mark and this rotation is then sensed by a quad detector that generates corresponding track error signals.

C3.4.3.4 Data coding and processing

An important step in storing digital data is encoding of the bit stream to be stored prior to recording, and, consequently, decoding of the readout signal after its conversion into digital form. Hereby, digital data is extracted from the analog signal obtained by the data detector. In a digital storage system, the input is typically a stream of binary data, i.e. binary digits 1 and 0, which has to be recorded onto a storage medium in such a way as to provide reliable and error-free recovery. The storage system is requested to record the data, to store it and to reconstruct it faithfully upon request. The digital data are first converted into analog signals and then stored as a stream of data marks correspondingly modulated and embedded into the recording format. For retrieval, the original digital data are extracted from the analog signal that is collected by the optical head via the data detector. From the user input to the recovered output, the data will undergo several steps of electronic processing including analog/digital conversion, equalization and filtering of the playback signal, error correction and modular coding, data synchronization and organization within the recording channel, etc [2].

Within the storage system, not only a recording unit is incorporated but also additional electronics units or subsystems that perform diverse steps in data processing from input to output after the stored data has been reconstructed by the optical read head. There is a variety of error sources that can cause errors in retrieval of information from a rapidly rotating optical disk. Some of them are related to the dynamic operation regime with the data structure reduced to the limit where the readout signal can only just be separated from the system noise. Also, media imperfections, defects, damages, etc lead to errors in reconstruction of information sequences stored at the affected locations. To retrieve the data faithfully, all sorts of error must be eliminated or compensated by appropriate error correction techniques. Therefore, the binary input data (user data) undergo several steps of encoding and modulation prior to being recorded on the storage medium.

The encoding process involves several measures against diverse error sources but it also entails other features that simplify the data processing and recovery. The flow of the data stream in an optical storage system is depicted in figure C3.4.14. Binary user data is subjected to two encoding processes prior to recording on the storage medium: error-correction coding (ECC) and modulation or recording coding. The encoding process includes typically one or more ECC steps followed by a modulation coding step where appropriate features are incorporated into the bit pattern. The ECC step is designed to protect the data against random and burst errors, and the modulation step organizes the data to be stored so as to maximize the storage density and reliability. In the ECC step additional bits, so-called check bits, are generated and added to the stream of user-data in order to create an appropriate level of redundancy in the overall bit sequence. Both the source data and the data emerging from the first (ECC) encoder are unconstrained, i.e. a randomly selected bit in the data stream may be either a '1' or a '0' with equal probability and arbitrarily long sequences of all 'ones' and 'zeros' may appear.

In addition to the error-correction, binary sequences are encoded by a modulation coding whereby the bit-patterns to be recorded are expanded by certain additional features. These enable the generation of a clocking signal for the electronic waveform and also provide more efficient usage of the storage area available on the disk. The modulation step involves mapping of small blocks from the error-correction coded sequence into larger blocks known as modulation code words. The data emerging after the modulation encoding step is usually d, k constrained, i.e. binary sequences are constrained in such a way that there must be at least d but no more that k zeros between two ones. Each binary segment consists therefore of a one that is followed by at least d but no more than k zeros. The data encoded in such a way is run-length-limited (RLL) and it is referred to as channel data because it becomes then converted into electrical waveforms used to control the recording process. The CD optical disk technologies rely on a block modulation code known as *eight-to-fourteen modulation* (EFM) that expands blocks containing eight data bits to 14-bit channel blocks. This modulation code is also applied in DVD systems in its advanced version, known as EFMPlus.

During retrieval, the run-length-limited channel data is reconstructed by the detector and processed by a modulation decoder which gives the error-correction coded binary data as output. The modulation

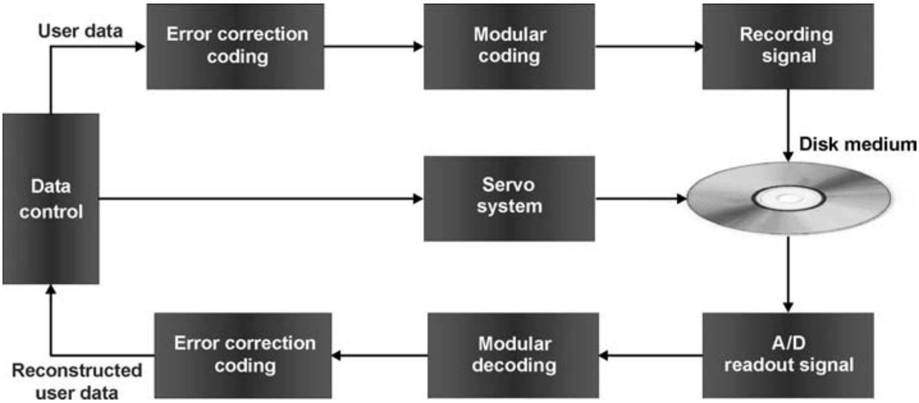


Figure C3.4.14. Channel data flow in an optical storage system.

decoder must correctly determine the logical ECC codeword boundaries within the stream of channel data to enable subsequent ECC decoding to take place. The ECC decoder processes the data leaving the demodulator to detect and correct any data error type for that it is designed to recognize and eliminate. Optical storage systems use typically *Reed Solomon* (*RS*) *codes* for error-correction coding. These are block codes, i.e. encoded data consist of codewords, or blocks, that contain a fixed number of bits. Encoding entails organizing the blocks of binary data into a succession of multibit symbols called information symbols, computing a number of additional symbols (of the same length) called parity checks and appending the parity checks to the information symbol to form a codeword. RS codes employed in optical data storage systems use eight-bit information symbols, i.e. they are designed to operate on bytes instead of bits. An RS code will correct up to a certain number of erroneous symbols in a given codeword, and for correcting a specific number of erroneous bytes, it needs twice as many parity check bytes. For example, one of the two ECC used in CD audio systems is the RS code which can correct up to 2 bytes of error in a 24-byte-long block of user data, with the addition of four parity bytes.

Information to be recorded on a disk is organized into uniformly sized blocks. Each of these blocks is written onto a portion of the storage medium that is referred to as a sector. Each track on the disk is then divided into a number of sectors that contain the user data and any ECC parity information related to it, i.e. calculated from it. Several different track formats find application in practical systems. As mentioned before, tracks can be realized either as concentric rings or a single continuous spiral whereby optical disks usually rely on the spiral format. For digital data storage, each track consists of sectors that are defined as small segments containing a single block of data. Spiral tracks are more suitable for writing of large data files without interruption while concentric rings better support multiple-operation mode when different operations such as write, erase, verify, etc are performed simultaneously in different tracks. The block of digital data stored within a sector has a fixed length which is usually either 512 or 1024 bytes. Each sector has its own address, the so-called header that specifies the storage location of a given sector on the disk.

The storage location is given by the track number and azimuthal position in the track at which the sector will be written. The information that constitutes a sector is usually written onto the medium in two parts. The first part of the sector is the sector header which consists of special patterns known as sector marks together with sector address data. The headers of all sectors are prerecorded, i.e. they are placed on the disk either when it is manufactured or when it is formatted for use. The additional space used by the codes and by the header within each sector constitutes the overhead, which may take between 10 and 30% of disk's raw capacity depending on application.

Prior to retrieving the data from the recorded storage medium, the optical detection head must access the medium and find the data that is requested by the storage system controller. The optical head must move to a particular radial position on the disk and it must find the track and sector that contain the requested data. Especially in the case of removable disk drives, the system checks other important information about the disk, for example, the sector size being used on the disk, the amplitude and polarity of signals obtained from prerecorded sectors and headers, etc. For reliable recording and readout, the design of a storage medium implements a defined recording format. The written data is embedded in this format which entails certain system information to be prerecorded at specific locations on the disk. The recording format also provides that the user data is written in its appropriate sector type at particular locations on the disk.

To conform to the recording format, the physical structure of the data to be stored on the disk must satisfy several requirements. In figure C3.4.15, two possible schemes for conversion of binary sequences into electrical waveforms are illustrated. In the so-called non-return-to-zero (NRZ) scheme, each bit is allotted one unit of time during which the voltage is either high or low, depending on whether the bit is 1 or 0. In the ideal case, recording with an NRZ waveform will result in identical marks that have the same length proportional to one channel bit time. Neighbouring data marks can have different

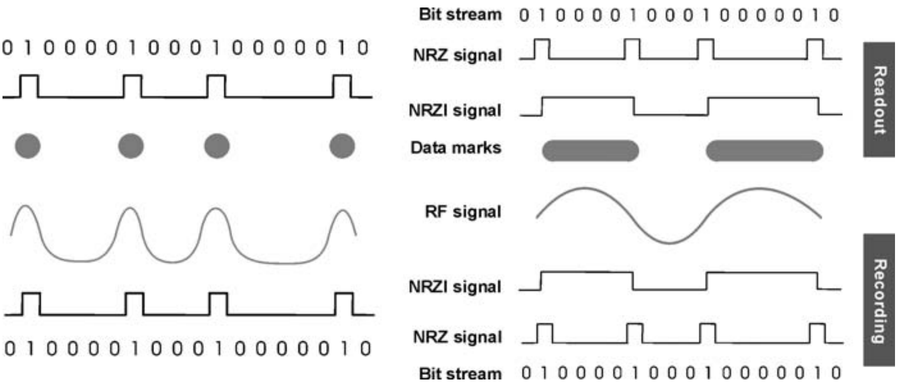


Figure C3.4.15. NRZ (left) and NRZI (right) conversion of binary sequences into electronic wavefronts.

centre-to-centre spacings whereby the centre of a mark represents a channel binary 'one'. The modified version of this scheme known as NRZI (non-return-to-zero-inverted) conforms much better to the EFM as here a 1 corresponds to a transition while a 0 is represented by no transitions at all. An NRZI waveform will induce data marks and intervening spaces that have variable discrete lengths. Here, the appearance of a binary 'one' corresponds to an edge of a recorded mark and both the mark length and the spacing between two successive marks are given by the number of binary 'zeros' between two ones. Using such a scheme, a single data mark can contain more bits than only one as in a simple coding scheme where each pit is allotted one bit.

C3.4.4 Novel approaches in optical storage

The acceleration of processor speeds and the evolution of new multimedia and Internet applications are creating an almost insatiable demand for high performance data storage solutions. Storage requirements are growing at an exponential rate encouraged by immense technological advances which have been achieved in recent decades. An ever-increasing amount of digital information is to be stored on-line, near-line and off-line combining magnetic, MO and optical systems. Rapidly growing demands are distributed through the data storage hierarchy, where diverse technologies are combined in complimentary way to satisfy specific requirements in different application environments. The usual computer applications will be served mainly by hard disk drives but emerging consumer applications that combine audio, video, 3D image, and computer data files are creating an important category with substantially different requirements. Optical disk technology has established itself as a mainstream product provider for audio, video and computer storage. The extraordinary success of recordable and rewritable disk formats based on CD and DVD technology has opened new prospects but also new requirements. Advanced storage of digital contents requires both higher storage capacity and fast data transfer. CD/DVD technology can satisfy immediate storage demands but novel application areas make essentially new technologies necessary.

In the future, optical data storage is expected to follow two directions to improve capacity and performance of disks that are currently available. The straightforward way predicts the further increase of the areal storage density by surpassing the limit imposed by the diffraction of light. Storage technologies that use only the surface of a medium for writing and reading are constrained to this direction. On the other hand, optical information storage uses laser-material interaction effects for recording and retrieval so that an entire spectrum of different optical phenomena can be applied to realize an optical memory. Developments in the field of nonlinear optical materials that exhibit strong

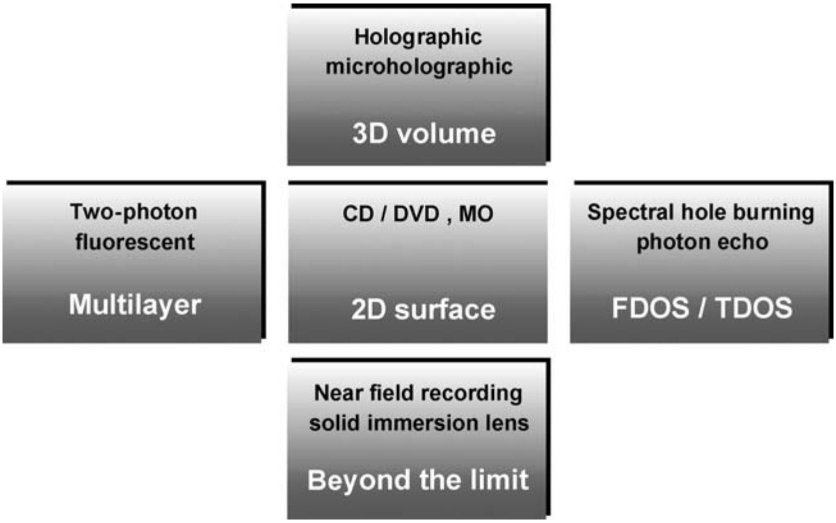


Figure C3.4.16. Novel approaches in optical storage to overcome the limitations of the surface storage by pits in CD/DVD as well as magneto-optical disk technology, including: optical storage beyond the resolution limit by near-field recording and solid immersion lens; frequency/time domain optical storage by spectral hole burning and photon echo memories; multilayer storage within transparent materials—two-photon and fluorescent memories; page-oriented holographic memories and bit-oriented microholographic disk.

laser-induced changes of their optical properties enable various novel approaches to become practically realizable. Using nonlinear optical effects, advanced technological solutions for optical storage may take advantage of new spatial and spectral dimensions (figure C3.4.16).

C3.4.4.1 Beyond the resolution limit

The traditional approach for increasing the areal density that has driven progress in data storage is to decrease the bit size. In optical storage, the attainable data density is largely determined by the size of the focused laser spot. A powerful way to surpass the density limit imposed by the diffraction of light in optical data storage is the usage of near-field optical recording. The creation and detection of pit marks smaller than that what is predicted by the diffraction barrier can be realized by numerous near-field optical techniques. For conventional optical systems, the achievable spot size governed by the diffraction law is $\propto \lambda/2\mathrm{NA}$. The resolution limit, known as the 'Abbe barrier', was empirically discovered and named for the German physicist, Dr Ernst Abbe, best known for his work in optics in the 1860s.

Recent progress in near-field optics has resulted in effective spot sizes smaller than 1/20 of the wavelength of light. To achieve such a high resolution, an aperture smaller than the resolution limit is placed between the light source and the medium. Light passing the aperture consists of propagating and evanescent waves; the smaller the aperture, the larger the fraction of evanescent field. The evanescent wave intensity decreases exponentially outside of the aperture; therefore, when the aperture-to-sample distance decreases, the evanescent power increases, and the resolution improves. If the aperture-to-medium distance is much less than the wavelength, the resolution will be determined by the aperture size rather than by the diffraction limit. Diverse techniques for near-field optical recording have been proposed making readout of sub-wavelength structures possible. However, a serious disadvantage for most of near-field recording (NFR) techniques using small aperture is low optical efficiency. Although near-field optical recording can provide extraordinarily high areal densities, it is difficult to satisfy requirements on high data transfer while maintaining a working spacing of less than a wavelength.

Scanning near-field optical microscopy (SNOM) makes it possible to overcome the diffraction limit of conventional far-field optical systems by placing the pick-up head very near (about 50 nm above) the media. In near-field microscopy, optical resolution beyond the diffraction limit is achieved by scanning a surface with the evanescent field behind a nanometre aperture. A small distance is necessary because the light field is confined only in the near field of the aperture. The technique can produce spots as small as 40 nm in diameter and conceptually can achieve areal densities in the order of 100 Gb in^{-2}. SNOM technology provides high areal densities but until now readout speed is low because the scanning process is very slow due to the small light power behind the aperture. Furthermore, the probe must be in near contact with the medium, making it difficult to prevent head crashes and support removable media. Another technique uses a metalized tapered optical fiber, in the end of which is a small aperture [3]. The tip of a fibre, which is smaller than the wavelength of the recording light, is positioned within 10 nm to the sample. This approach has been already used to write and detect 60 nm diameter MO domains. However, the tapered fibre approach also suffers from very low optical efficiency.

A technique that might solve the trade-off between an extremely high resolution and practical system implementation is based on the solid immersion lens (SIL). The principle of the SIL is that by focusing light inside a high refractive index glass where the propagation speed is slow, the spot size can be reduced below the minimum achievable spot size in air [4, 5]. The SIL reduces the actual spot size by both refracting the light rays at the sphere surface and by having an increased index of refraction within the lens. The hemispherical glass of refractive index n receives the rays of light at normal incidence to its surface (figure C3.4.17). These are focused at the centre of the hemisphere to form a diffraction-limited spot that is smaller by a factor n compared to what would have been in absence of the SIL. That becomes obvious if we consider the minimum achievable spot size which is given by

$$d_{\min} = 0.61\frac{\lambda}{n\sin\theta}$$

where θ is the aperture angle of the focusing lens. The solid immersion lens allows the aperture angle to be increased also. An increase of numerical aperture from 0.6 as typical for conventional optical disk systems to 0.95, and a refractive index change from 1.0 to 2.2 would result in a spot size of 0.2 μm at the flat surface of a solid immersion lens for light at 670 nm.

Although SILs cannot produce spot sizes as small as tip fibre apertures can, they have the advantage of a substantially higher optical throughput. Another advantage is that an SIL can be with ease integrated in any conventional system configuration as an addition to the objective lens. The application of the SIL also requires an extremely short working distance of the lens to the recording

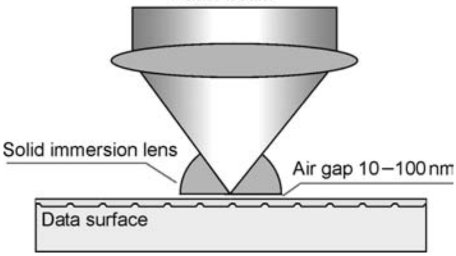


Figure C3.4.17. Solid immersion lens for near-field optical storage. The light is focused internally in a semi-spherical lens onto the flat surface.

layer—about 100 nm—as well as a very thin protective layer. Flying heads as in hard disk storage can be used to provide such a distance, which is maintained over a disk that rotates rapidly. The SIL combined with short wavelength lasers should enable the technique to reach areal densities of more than 10 Gb in^{-2}. To date, implementations of near-field recording with SIL are planed in MO storage. The first NFR products announced will have a storage capacity of 20 GB per disk.

C3.4.4.2 Multidimensional optical information storage

Looking for solutions to overcome the restrictions of two-dimensional optical storage systems, various approaches have been explored that use one further dimension in spatial, spectral or time domain. Novel technologies, such as holographic storage, two-photon or fluorescent memories, persistent spectral hole burning, photon echo memories, etc, are at various stages of development. Opening a new dimension in addition to the two-dimensional surface of a storage medium, they have the potential to improve tremendously both capacity and data transfer rates of optical storage systems.

In present optical storage systems, one-dimensional serial information is stored in a two-dimensional medium. In three-dimensional optical memories, three independent coordinates specify the location of information. Three-dimensional optical information storage has been proposed first in 1963. One of the unique characteristics of optical volume storage is the very high bit packing density that can be attained. The ultimate upper limit of the storage density in three-dimensional storage media is of the order of $1/\lambda^3$, compared with $1/\lambda^2$ for surface or two-dimensional optical storage media. This results in 10^{12} to 10^{13} bits cm^{-3}, although the practical limit set by other parameters of the optical system and the constraints of the recording material may be lower than this.

Three-dimensional optical storage systems may generally be classified as bit-oriented and page-oriented. In holographic page-oriented memories the information associated with stored bits is distributed throughout the whole volume of the storage medium. In bit-oriented memories each bit occupies a specific location in three-dimensional space. Various approaches to realize a three-dimensional optical memory by bit-oriented storage have already been presented not only including holographic but also non-holographic. Such storage methods are based in general on a point-like or bit-by-bit three-dimensional recording by creating small data marks within the medium. Hereby, data marks represent single bits and are defined by large contrast in some optical property of the storage material. Using a nonlinear optical response of the material, the optical interaction can be confined to a focal volume. With a diffraction limited, focused light beam, the physical size of bits can become as small as the wavelength of the laser beam in all three dimensions. Owing to the submicrometre dimensions of a single bit, bit-oriented optical data storage requires very strict tolerances for the focusing optics and recording alignment.

Optical storage in form of holographic volume gratings has been investigated during the past three decades as a straightforward approach to realize three-dimensional high density memories. In addition, alternative solutions for three-dimensional optical memories are also under investigation. These include the extension of present disk systems based on phase-change or MO media to a layered format but also various new concepts of multilayer optical memories. In this case, the third dimension is introduced by recording the data in multiple layers through the thickness of a volumetric storage medium.

C3.4.4.3 Multilayer optical information storage

The simplest way to use the third dimension of a storage medium is multilayer storage. Using multiple data layers instead of one, the overall storage capacity will grow linearly with the number of layers. Data layers are separated by thin transparent spacers and addressed separately by a highly focused laser beam. The success of dual-layer DVD-ROM has attracted interest but layered optical storage as natural extension of conventional optical disks based on readout from the reflective layer has only a moderate potential to increase the storage capacity. The number of layers per side of a disk is limited strongly by higher optical power requirements, interlayer cross-talk and aberrations that appear while focusing to several layers at different depths simultaneously. Combined with other recording techniques, multilayer approach can become more attractive. In the case of two-photon or fluorescent memories that use transparent materials as storage media, the number of layers can become very large. Such quasi-three-dimensional optical memories use the volume of a storage medium by recording the data as binary planes stacked in three dimensions. The data is stored by discrete bits in the plane, but also through the volume.

Following the experimental advances in media and system concepts made in the last years, optical recording by two-photon excitation in photochromic as well as photorefractive media became very attractive as an alternative for three-dimensional optical memories. The modulation of the recording material is usually localized by using focused Gaussian beams. Two-photon excitation refers to the simultaneous absorption of two photons, whereby the excitation rate for this process is proportional to the square of the writing light intensity. Therefore, the excitation remains confined to the focal volume corresponding to the intensity distribution of a focused Gaussian beam. The basis of a two-photon recording system is the simultaneous absorption of two photons whose combined energy is equal to the energy difference between initial and final states of the recording material. This simultaneous absorption results in a structural phase transition that is reversible and detectable by measuring the fluorescence of the material. The read beam is unabsorbed and passes through the unwritten areas of the material while the recorded data marks will cause the absorption of the readout beam exciting the fluorescence at a longer wavelength.

A variety of materials have been proposed for two-photon recording. The most important material requirements concern the photochromism, i.e. the ability to change the chemical structure under light excitation, a fluorescence in one of two chemical states, stability of both states at room temperature, etc. A typical system involves two beams that are called data beam and address beam as shown in figure C3.4.18 . The data beam at 532 nm is modulated with a spatial light modulator (SLM) and focused at a particular plane within the medium. The addressing beam at 1064 nm provides the second photon required for the two-photon excitation. The data are written in the overlap region of the two beams and then read out by fluorescence when excited by single photons absorbed within the written bit volume. Hereby, the data beam is blocked and the read beam at 532 nm is focused to reconstruct the selected data

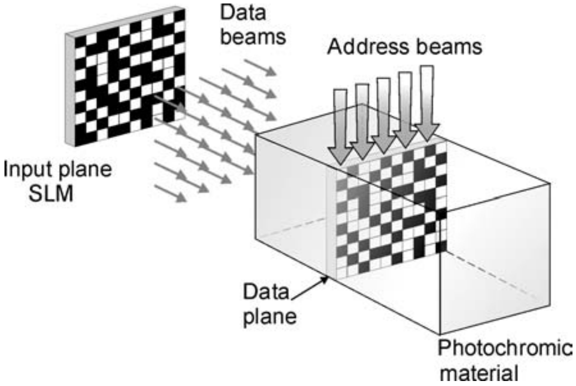


Figure C3.4.18. Three-dimensional optical storage by multilayer recording. Two-dimensional binary data planes are stacked in the depth of a transparent storage medium. Readout relies on laser-excited fluorescence.

page within the volume. The readout plane is then imaged onto a CCD. The spot size is limited by the recording wavelengths through diffraction. The approach promises not only ultra-high effective areal density but also parallel access to the stored data [6, 7].

C3.4.4.4 Frequency/time domain optical storage

Frequency or time domain optical storage techniques adhere to two-dimensional media but open one additional dimension in frequency or time domain. Persistent spectral hole burning (PSHB) takes a step ahead of one-bit-per-spot memories, allowing multiple bits to be written, erased and rewritten in a single location. High densities become possible because the diffraction limit does not limit PSHB memories to the recording of a single bit per spot, as it does in conventional optical data storage. In a PSHB material, it is possible to discriminate many different spectral addresses within a single λ^3 volume.

Frequency domain optical storage based on PSHB involves burning ‘holes’ in a material’s absorption band [8, 9]. The ideal material has many narrow, individual absorption lines that form a broad absorption band. Practical recording in PSHB media may be done as an extension of conventional holography with the difference that, instead of a single wavelength, a large number of independent spectral channels may be used. This number ranges from 10^4 to more than 10^7, depending on the material. For writing information, a frequency-tunable laser is focused on a single spot scanning down in wavelength to induce transition from one stable state to another in an absorbing centre. As a result, there will be holes at certain frequencies that correspond to the presence of the written bits. Persistent spectral hole burning using the frequency domain promises storage densities up to $10^3\,\mathrm{bits}\,\mu\mathrm{m}^{-2}$ which could be achieved with 10^3 absorbing centres, i.e. spectral holes per diffraction limited laser spot.

Another PSHB storage method, so-called photon echo optical memories, is based on time-domain storage [10, 11]. Time-domain hole burning also uses spectral holes for data storage, but relies on coherent optical transient phenomena. Time-domain storage is realized by illuminating an inhomogenously broadened material by two temporally separated resonant optical pulses. The first (reference) pulse creates an optical coherence in the material. The second pulse is temporally encoded with data and interferes with the optical coherence created by the reference pulse resulting in a frequency-dependent population grating. The recorded information is read out by illuminating the storage material with a read pulse (identical to the reference pulse) which generates a coherent optical signal having the same temporal profile as the data pulse.

The maximum number of bits that can be stored in a single spot using PSHB is given by the ratio of the inhomogeneous $\Delta\omega_{\mathrm{i}}$, and homogenous $\Delta\omega_{\mathrm{h}}$, absorption line width of the storage material. This ratio can range up to 10^7 in some materials. Because of the large number of frequency channels available, high storage densities may be possible with reasonable laser spot sizes. Even though both of these methods have advantage of increased storage density ($\geq$Mbit/laser spot), their application capability is significantly limited by the operating temperatures which should be kept extremely low around liquid helium temperature. In particular, present research efforts concentrate on achieving room temperature hole burning with novel materials, but there are still a number of technical challenges to overcome before PSHB becomes viable for data storage.

C3.4.5 Holographic information storage

In contrast to three-dimensional multilayered optical memories, in holographic storage, the information is recorded through volume. Recording is accomplished by interfering two coherent laser beams, the information-bearing signal beam and the reference or address beam. The resulting intensity pattern is then stored in a photosensitive material by inducing a grating-like modulation of its optical properties such as refractive index or absorption coefficient. The data is reconstructed by diffracting the address beam at the induced grating. A unique characteristics of thick holographic gratings is the Bragg-selectivity which allows many holograms to be stored overlapping by applying appropriate multiplexing methods.

C3.4.5.1 Page-oriented holographic memories

Holographic memories usually store and recall the data in page-format, i.e. as two-dimensional bit arrays which offers the way to realize high data rates and fast access. Combined with multiplexing, the inherent parallelism of holographic storage can provide a huge increase in both capacity and speed. For more than 30 years, holography has been considered as a storage approach that can change standards and prospects for optical storage media in a revolutionary manner. Depending on a number of supporting technologies, holographic memories became realizable with advances in photonics technology, particularly with improvements in liquid crystal modulators, charge coupled devices, semiconductor detectors and laser sources. Ongoing research efforts have led to impressive advances [12]. The first completed working platforms demonstrated high storage densities of more than 300 bits μm^{-2}, but they are still far from commercialization.

In contrast to the conventional optical recording where an individual data bit is stored as localized change in some optical property of two-dimensional storage media, holographic recording allows to store the data page-wise in the volume of the material. Instead of storing one single bit at each location, large data pages can be recorded and read out at once. The information to be stored is first digitized, then loaded onto a spatial light modulator as a two-dimensional pattern of binary ones and zeros. The SLM imprints that binary data page to the signal beam. The data is recorded by intersecting the signal beam with a reference beam inside the storage medium (figure C3.4.19). The three-dimensional interference pattern induces a corresponding spatial modulation of the refractive index of the recording

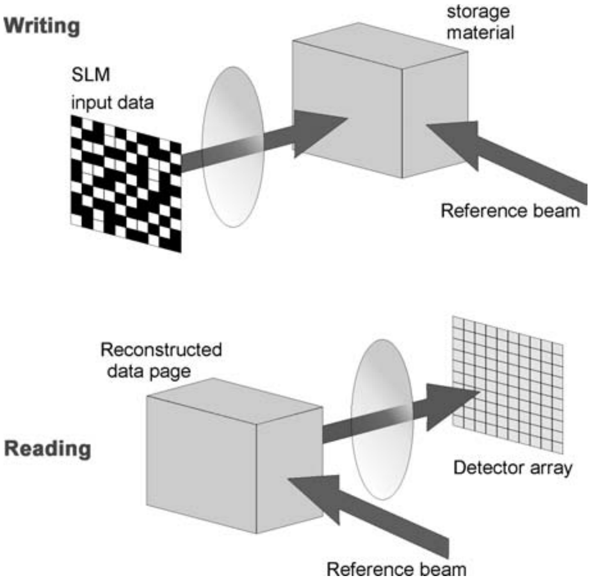


Figure C3.4.19. Page-oriented holographic storage. A two-dimensional data pattern created by spatial light modulator is stored by interfering signal and reference beam to record the hologram. One whole page is written at once, many pages overlap in the same volume. One page from many is read out with the corresponding reference beam.

material. Consequently, the data is stored as a refractive index grating representing a phase volume hologram.

As in page-oriented holographic storage each data bit is distributed in three dimensions through the recording medium, there is no direct correlation between the data bit and a single volume element in the recorded structure. The stored data are retrieved by introducing the same reference beam used to record it, and read out by projecting the reconstructed signal beam onto the output detector array for optic-to-electronic conversion.

Due to the Bragg-selectivity of holographic volume gratings, it is possible to store several holograms overlapping in the same volume element by either changing the angle or the wavelength of the reference beam. Such overlapping recording of multiple holograms in one single position is referred to as holographic multiplexing. In a multiplexing scheme, holographic structures are mixed together whereby the Bragg effect allows retrieval of an individual data page while minimizing cross-talk from other pages stored in the same volume. The diffraction efficiency varies according to the mismatches in angle or wavelength between recording and readout. Deviations from the Bragg condition lead to a rapid decrease in diffraction efficiency which makes a selective reconstruction of multiple holographic gratings possible.

Various multiplexing methods have been proposed in diverse holographic storage systems. Wavelength and angle multiplexing result directly from the Bragg-selective character of thick volume holograms. The addressing mechanism here is the wavelength or angle of incidence of the reference beam. Both methods allow many holograms to be recorded in the same position but their practical impact is limited. Angle multiplexing requires complex optics, and the crucial component for efficient wavelength multiplexing is a laser light source that should be tunable in a sufficiently wide range. Such lasers are available but complex so that a significant increase of storage density by wavelength multiplexing only is difficult to realize in practical systems. New methods such as peristrophic or shift multiplexing have been developed for holographic storage systems which use a disk-shaped medium instead of a photorefractive crystal cube. *Peristrophic multiplexing* is based on the rotation of the plane-wave reference beam around the optical axis; during the readout, the reconstructed holograms follow the motion of the reference beam. The address of an individual hologram is the relative rotational position of the storage medium. *Shift multiplexing* relies on a spherical wave as the reference beam while the signal beam is still a plane wave. Such spherical holograms can be shifted relative to the reference and stored next to another with a shift distance below 10 μm. The relative displacements are small enough that holograms in subsequent locations overlap significantly.

Volume holography is a powerful approach for digital storage systems with high densities (≥ 100 bits μm^{-2}) and fast transfer rate ($\geq$ Gbit s^{-1}). However, the practical realization of holographic memories suffers from the lack of suitable storage media. Both the performance and viability of systems under development are significantly limited by the characteristics of the available materials. To date, the requirement of adequate storage materials has been one of the most crucial aspects in development of holographic memories. Indeed, there are very rigorous demands on storage materials which should be satisfied to realize holographic memories as competitive, reliable optical storage systems of improved performance. The search for an optimum material to be used in holographic data storage does not appear to be finished yet. An ideal recording material should be of high optical quality, it should be able to hold the recorded data for a long time, and, for commercial applications, very reliable and not too expensive. Considering the physical processes, one can define a number of parameters to be controlled. The most important for the viability of holographic storage are high resolution (>3000 lines mm^{-1}), high photosensitivity, large dynamic range (i.e. diffraction efficiency of multiplexed data pages), archival storage time, low absorption, low scatter. In particular, a large dynamic range of the storage material is necessary to allow multiplex recording of many holograms in the same volume.

Many kinds of materials have been investigated as holographic storage media. With sufficient material development efforts, the necessary optical quality has been achieved for both inorganic photorefractive crystals and organic photopolymer media. Photorefractive crystals such as lithium niobate, barium titanate and strontium barium titanate, were used previously in holographic systems. In the last years, a new class of photosensitive polymers has been introduced to satisfy the demand on adequate materials for holographic storage. Depending on the recording material, different optical system architectures have been developed for holographic memories. Holographic storage media can be classified in two categories: thin photosensitive organic media and thick inorganic photorefractive crystals. Thin media (a few hundreds micrometres thick) are most suitable for transmission type architecture using a variety of shift or phase multiplexing techniques, while angular multiplexing in various modifications is usually applied in thick media (about centimetre thick). A typical system architecture in this case is based on the 90° geometry while thin photopolymer layers are often used in a disk-based configuration. In this concept, digital holographic pages are stored on a disk-shaped medium and organized in tracks similar to those on conventional optical disks. The disk can rotate continuously and it can also move across tracks to allow the optical head to access the entire area of the medium. The storage medium is typically an organic photopolymer layer sandwiched between two glass plates.

C3.4.5.2 Bit-oriented holographic storage

An alternative to page-oriented holographic memories is three-dimensional bit-oriented optical storage on a holographic disk medium (figure C3.4.20). The so-called microholographic approach offers a compromise by combining the bit-wise storage of CD/DVD and holographic volume recording, which makes it possible to advance the capabilities of conventional disk technologies by implementing spatial and wavelength multiplexing [13, 14]. Microholography expands surface storage into three dimensions by storing the data as microscopic volume gratings instead of pits. Holographic recording is realized within a system that is in its main features very similar to the recordable CD or DVD systems. The technology provides volume storage of information while the areal structure of the stored data remains comparable with pits on a DVD. The microholographic storage concept benefits from both technologies: The bit-oriented storage allows for using many solutions of the highly developed CD/DVD technology. In addition, holographic recording offers a path to overcome the limitations of this technology which are related to its two-dimensional nature.

On a microholographic disk the pit-land structure of a CD is replaced by localized volume gratings which allow the bit-oriented storage to use the third dimension. The storage medium is a thin photopolymer layer coated onto an optical disk substrate. For writing, microscopic reflection gratings are holographically induced to vary the reflectivity of the disk locally. A laser beam is focused into the photosensitive layer and reflected back with the mirror. The interference pattern of the incident and reflected beam induces a grating-like modulation of the refractive index of the storage medium.

To retrieve the stored data, the original signal beam is reconstructed by reflection of the read beam at the induced gratings. Recording with sharply focused laser beams results in localized volume storage. The microgratings can be packed very densely and arranged in tracks similar to those on a CD. When the disk is rotating, microgratings of variable length are induced dynamically whereby grating fringes are extended in the motion direction.

The storage system is very similar to the conventional recordable or rewritable optical disk systems: A focused laser beam is used for writing and reading, the data is stored bit-wise in tracks on the rotating disk, and similar systems for automatic focusing and track following are needed to control the position of the laser beam focus on the rotating disk. All these common aspects simplify the practical realization of the microholographic system as many components developed for the CD/DVD technology can be directly used or adapted for this purpose. The main difference here is the reflecting unit underneath

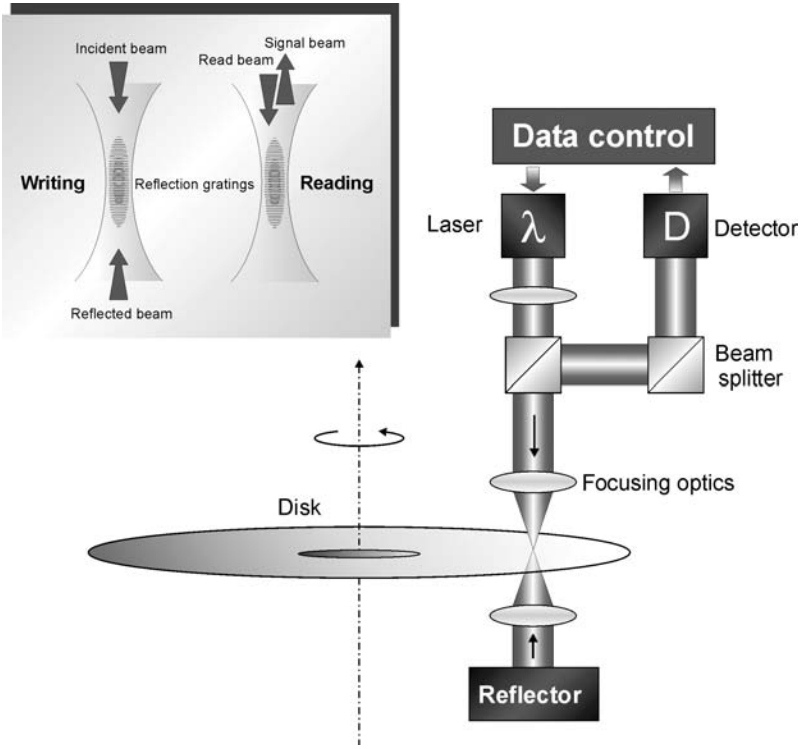


Figure C3.4.20. Microholographic storage system. Microgratings are created in localized volume elements by sharply focused laser beams. The stored data is read out by detecting the local reflectivity of the disk; the reflecting unit is removed in this case. High storage densities can be achieved by combining wavelength multiplexing and multilayer storage.

the disk that is needed for writing. During readout, this unit, in the simplest realization an aspherical mirror, can be tilted or removed so that only reflection from the gratings will be detected.

The physical structure of the stored data is similar to the pit-land structure on a CD/DVD as stripe-shaped microgratings are written dynamically while the grating length corresponds to a coded bit-stream. Multiplexing is integrated parallel into the coding scheme in order to maximize writing and reading speed. Each wavelength/layer channel corresponds to standard data channels in CD/DVD systems so that recording and error correction code algorithms evolved in conventional technologies can be directly used. This makes it possible to satisfy the requirement for downward compatibility with CD/DVD media.

While the microholographic disk design is very simple (only one photopolymer layer on a substrate), the storage density can still be increased by writing data in multiple layers. To address the layers, the beams are focused to different depth inside the photopolymer layer. The main advantage of the microholographic approach is that it takes bit-oriented storage to the third dimension: by using holographic multiplexing methods tracks can be overlapped in the same volume. This way very high storage densities of more than 100 bits μm^{-2} can be achieved.

Application of wavelength multiplexing allows a linear increase of the storage density and also of the write/read rate with the number of wavelength used. Wavelength multiplexing is realized by simultaneous recording of several gratings in the same position with write beams of different wavelengths. In this case, a complex periodical grating structure is induced that contains all single-colour gratings. Due to the wavelength selectivity, each laser beam with a certain wavelength detects only the corresponding grating during readout. The data rate is increased since all bits contained in one multiplex grating can be recorded and read in parallel. In addition, localized volume storage allows for spatial multiplexing which is realized as multilayer storage. Microgratings are recorded in different independent planes within the storage medium. A single data layer is addressed by the focused laser beam. Combining the wavelength multiplexing and multilayer storage, high storage densities can be achieved.

In microholographic storage, strong holographic requirements for an adequate material are relaxed by the dynamic bit-by-bit recording while at the same time multiplexing improves performance significantly. Furthermore, manufacturing technology can provide low-cost, removable media and downward compatible systems. The microholographic disks may be successful as removable storage media that satisfy high-capacity demands in specific areas including data banks, archiving and security systems, image processing and multimedia applications.

C3.4.6 Potential impact of novel technologies

Optical systems for recording and retrieval of digital information represent a rapidly developing field with a huge potential to encompass entirely new applications and to provide solutions to problems arising from these applications. Continuous advances and discoveries in related technologies, devices, and materials have opened an entire spectrum of optical effects and materials that can be used to provide writing and reading mechanisms in an optical data storage system.

Trends towards rewritability and higher storage capacity have moved optical storage into competition with high-end magnetic storage. The key difference between these two technologies is the removability of optical media but also their excellent robustness, archival lifetime and very low cost. An additional advantage of optical technology is the stability of written data, a feature that makes optical media suitable for archival lifetimes. An optical disk can be removed after recording and read out in any compatible drive which enables data to be stored separately from the main computer system or network. Typical applications range from archival storage, including software distribution, digital photographs and imaging, movies and other video materials.

For archival storage many disks are organized in an optical library system capable of storing and managing large amounts of data. Library systems are usually constructed as jukeboxes comprising hundreds of disks to provide high capacities for long-term storage. An important advantage of optical archiving systems is that the data is stored off-line which releases computers or networks but also provides data security and retrieval even if the network is irreparably destroyed. Optical systems are already widely accepted in enterprise and institutional storage with applications ranging from extending existing server capacities to publishing and image storage. One of the most important applications for optical archival storage is document and image management—where documents and images such as receipts, x-rays, photographs and other records are stored in digital form on optical disks. Plurality of disks are then arranged in a database to facilitate rapid retrieval.

Optical storage media presently available are CD, DVD and MO disks. CD and DVD rely on identical recording and readout processes while the differences originate from the format specifications. Despite impressive advances and continuous increases in storage density, the existing bit-wise, serial-access optical storage is far away from realizing the full potential of optical technology. With opening computer applications such as three-dimensional imaging, video mail and server applications including, for example, digital libraries, satellite imagery, medical document and image archiving, optical storage should ensure capacities exceeding a terabyte per disk. Numerous techniques provide the ability to achieve high storage capacities by a more effective use of the volume of a storage medium and/or by taking advantage of additional degrees of freedom such as the recording wavelength. Moreover, the inherent parallelism of optics offers the possibility to record and retrieve large data files with data rates exceeding a gigabyte per second.

Volume holographic storage as well as two-photon or fluorescent storage hold promises for high-capacity, high-speed systems. In addition, microholographic disks or fluorescent multilayer disks that store the data bit-wise as ‘fluorescent pits’, can also satisfy the requirements for downward compatibility and low-cost media. Figure C3.4.21 gives a comparison of storage densities achievable in different technologies. A crucial aspect for the reliability of all these systems is the storage material itself. Many types of materials have been investigated in recent years as optical storage media including inorganic photorefractive crystals, organic photopolymers and biological systems such as protein bacteriorhodopsin or DNA polymers. Progress in the last couple of years has been impressive, particularly in the field of photosensitive polymers that offer a wide variety of possible recording mechanisms including both write-once and rewritable media. In particular, new photopolymer materials have been introduced for holographic storage. Optimization and further development of photopolymer media will be the key to the success of this and other advanced optical storage technologies.

Page-oriented holographic memories hold the top of the table, promising terabyte devices and Gbit/s data rates, but it is questionable if they would be able to compete with the existing optical disks in daily life. Holographic storage can find many specific application, such as data banks, where large data files have to be stored and recalled with fast access. The possibility of associative retrieval enables holographic memories to be used as data search engines, i.e. content-addressable database servers or large Web servers.

In small end-user systems, the requirements will rather be governed by the convergence of entertainment and computing. From this point of view, bit-oriented optical storage offers more realistic

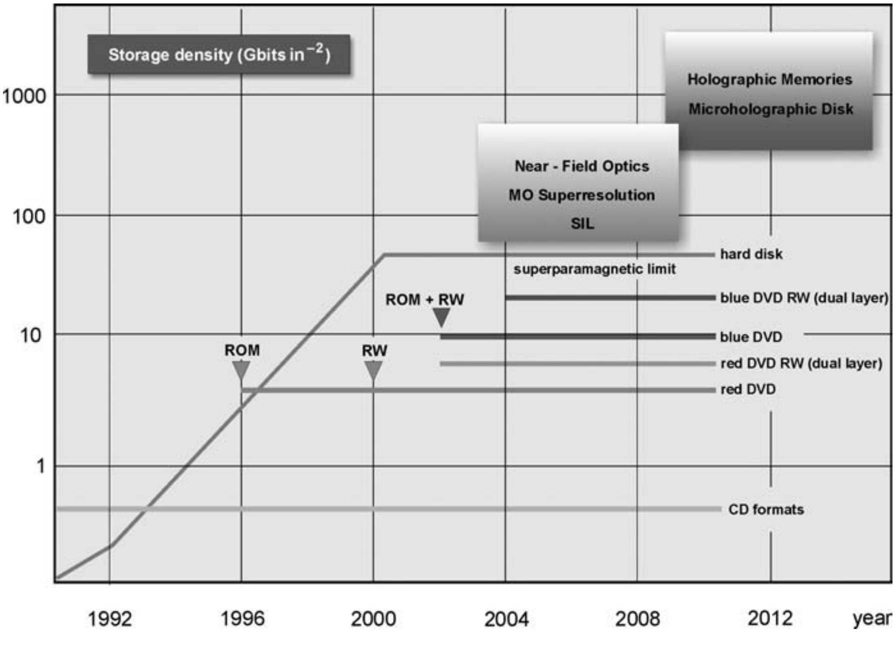


Figure C3.4.21. Potential impact of novel optical technologies in comparison with achievable areal densities in conventional optical storage systems and hard disk systems.

solutions for the next generation. With the availability of adequate storage media, the microholographic disk technology might provide the successor to the blue DVD generation. The recording technique can improve performance significantly, reaching with ease the threshold of 100 GB per disk. Such advanced optical disks are asked for many applications and are particularly attractive for three-dimensional imaging and video storage.

Although recent advances towards commercial devices are impressive, it remains to be seen which technology will be successful in providing the next generation optical storage media. A wide variety of materials as well as recording and readout techniques are under development. The requirements of foremost importance in optical information storage are high performance in terms of capacity and data rates, removability and compatibility. The demand on downward compatibility is to ensure the use of universal drives, but also reliability and low-cost media. Meeting these requirements simultaneously is a major challenge to developers of novel optical technologies. Competition combined with an immense market potential and appetite for storage keeps the field exciting.

References

[1] Mansuripur M 1995 *The Physical Principles of Magneto-Optical Recording* (Cambridge: Cambridge University Press)
[2] McDaniel T W and Victora R H 1997 *Handbook of Magneto-optical Data Recording* (New Jersey: Noyes Publications)
[3] Betzig E, Trautman J K, Harris T D, Weiner J S and Kostelak R L 1991 *Science* **251** 1468
[4] Mansfield S M and Kino G S 1990 *Appl. Phys. Lett.* **57** 2615
[5] Terris B D, Mamin H J and Rugar D 1996 *Appl. Phys. Lett.* **68** 141
[6] Parthenopolous D A and Rentzepis P M 1989 *Science* **245** 848
[7] McCormick F B, Zhang H, Walker E P, Chapman C, Kim N, Costa J M, Esener S and Rentzepis P M 1999 *Proc. SPIE* **3802** 173
[8] Moerner W E 1985 *J. Mol. Electron.* **1** 55
[9] Caro C D, Renn A and Wild U P 1991 *Appl. Opt.* **30** 2890
[10] Mossberg T W 1982 *Opt. Lett.* **7** 77
[11] Bai Y S and Kachru R 1993 *Opt. Lett.* **18** 1189
[12] Coufal H J, Psaltis D and Sincerbox G T 2000 *Holographic Data Storage* (New York: Springer)
[13] Eichler H J, Kuemmel P, Orlic S and Wappelt A 1998 *J. Sel. Top. Quantum Electron* **4** 840
[14] Orlic S, Ulm S and Eichler H J 2001 *J. Opt. A: Pure Appl. Opt.* **3** 72–81

C3.5

Optical information processing

John N Lee

C3.5.1 Introduction and historical background

Optical technology has been developed for highly effective transport of information, either as very high speed temporal streams, e.g. in optical fibres or in free-space, or as in high-frame-rate two-dimensional (2D) image displays. There is, therefore, interest in performing routing, signal-processing and computing functions directly on such optical data streams. The development of various optical modulation, display, and storage techniques allows the investigation of processing concepts. The attraction of optical processing techniques is the promise for parallel routing and processing of data in the multiple dimensions of space, time, and wavelength at possible optical data rates. For example, in the temporal domain a 1 nm wide optical band at a wavelength of 1500 nm has a bandwidth of approximately 100 GHz, and temporal light modulators with 100 GHz bandwidth have also been demonstrated for optical fibre systems (chapter B5) [1, 2]. However, notional optical processing techniques can be envisioned that handle many such narrow-wavelength bands in parallel, and also operate in a combined spatio-temporal domain. Employing all domains simultaneously, it is theoretically possible to perform spatio-temporal routing and processing at an enormously high throughput in the four dimensions x, y, t, and λ. Throughput of 10^{14} samples/s would result from simple examples based on feasible modulation and display capabilities. In one case a 100 GHz temporal modulators can be combined with wavelength-selective devices to provide several hundred 1 nm wide channels at the operating wavelengths of existing photodetectors and light sources. A second example would consider that 2D spatial light modulator (SLM) devices can be constructed to have $>10^7$ pixels/frame (see chapter C2.3) and that material developments allow optical frame update rates on the order of 1 MHz (chapter B14) [3]. Unfortunately, although an optical processor operates on data in optical form, it is presently not possible to equate these maximal modulation and display rates to the expected information-processing throughput rates of such processors. There are penalties on the throughput due to necessary data pre-processing and post-processing in any information-processing system. These include the need to format and condition the input data to a processor, to compensate for shortcomings of any analogue signals (e.g. nonuniformities in space and time), and perhaps most importantly, to examine the processor's output data and extract the useful information. The latter is often an iterative process and requires fusion with other data processing results. An optical processor's speed advantage could be largely negated unless all processing operations can be performed at speeds commensurate with modulation and display rates. Thus, equally important considerations are the need to identify those operations that can be effectively performed optically, and the need to develop optical processing architectures that minimize the penalties on optical throughput. Because of these considerations optical information-processing approaches have covered a wide range of topics.

Therefore, we first provide a brief review of the various paradigms that have been investigated in optical processing.

C3.5.1.1 Analogue optical processing

In the oldest paradigm, optical data in analogue form can be manipulated to perform useful functions. The classic implementation of an analogue spatio-temporal processing function is the use of a simple lens to produce, at the back focal plane of a lens, the complex Fourier transform (both phase and amplitude) of optical data at the front focal plane [4]. In the most common and simplest configuration, shown in figure C3.5.1 a one-dimensional (1D) object with complex transmission,

$$t(x_0) = a(x_0)\exp[-j2\pi b(x_0)] \tag{C3.5.1}$$

where a and b are the amplitude and phase value at pixel location x_0, is positioned at a distance d in front of a lens of focal length f_1. Illuminating the object with coherent light of wavelength λ, one obtains at one focal length distance behind the lens the amplitude distribution [4]

$$U_f(x_f) = \frac{A\exp[(j\pi/\lambda f_1)(1 - d/f_1)x_f^2]}{j\lambda f_1}\int_0^L t(x_0)\exp[-j(2\pi/(\lambda f_1))(x_0 x_f)]dx \tag{C3.5.2}$$

where L is the spatial extent of $t(x_0)$, A is a constant, and the subscripts 0 and f are used to denote the object and output space, respectively. When

$$f_1 = d \tag{C3.5.3}$$

aside from the finite spatial limits of the integral, a Fourier transform of $t(x_0)$ results (to within a constant multiplicative factor). The 1D example given in equation (C3.5.2) employs a cylindrical lens. Use of a spherical lens results in a 2D Fourier transform, but since the x and y variables in the exponential term of the Fourier transform are separable, a spherical lens can be used even for the 1D data in figure C3.5.1. Alternative lens configurations can produce the Fourier transform but require two instead of one lens, or produce the correct amplitude of the Fourier transform but with a curved phase front [4].

The core idea of Fourier transformation by a lens is the basis for many demonstrations of single-function optical processors. Electro-optic and acousto-optic modulation devices have been developed

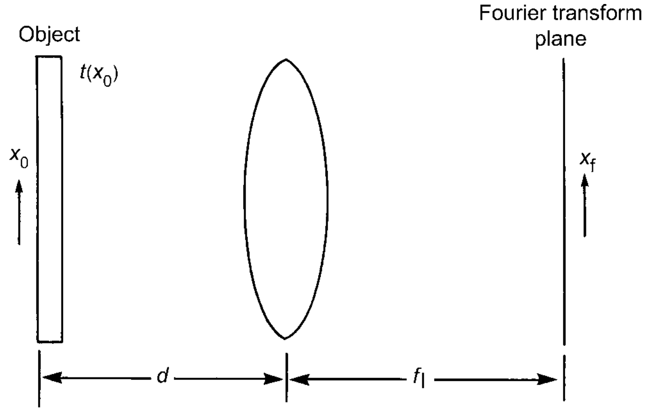


Figure C3.5.1. Optical arrangement for Fourier transformation with a single lens.

for optical input and output at the focal planes. To exploit this powerful concept, other mathematical integral transforms have often been re-cast as Fourier transformation problems, in particular the matched filter operation, or equivalently the correlation integral

$$R(\tau) = \int s_1(\mathrm{t}) s_2^*(t+\tau)\mathrm{d}t. \tag{C3.5.4a}$$

According to the convolution theorem, equation (C3.5.4*a*) is mathematically equivalent to the Fourier transform of the product of the instantaneous Fourier transforms of the two signals, $S_1(\omega)$ and $S_2{}^*(\omega)$, i.e.

$$R(\tau) = \int S_1(\omega) S_2^*(\omega)\exp(\mathrm{j}\omega\tau)\mathrm{d}\omega. \tag{C3.5.4b}$$

Alternative transform kernels to the Fourier Transform have been developed, and implemented using computer-generated holograms rather than lenses [5]. Processing applications involving massive amounts of linear processing have been successfully addressed with prototype analogue signal-processing systems that perform temporal spectral analysis [6] and correlation [7, 8]. However, the optical processing performed in these systems implementations are linear operations involving only multiplication and addition. These analogue processors therefore address an application niche in signal processing, and because they are analogue, must contend with accuracy and dynamic range concerns. Single-purpose analogue computation engines, described in the references, have been embedded into conventional electronic processing systems to accelerate specific signal-processing tasks, but not to perform general computations. It has often been the case that insertion of such engines leads to bottlenecks due to the optical-to-electronic and electronic-to-optical conversions, and the incompatibility of optical processor speeds with electronic computer limits such as memory access time. An approach that potentially relieves these bottlenecks involves ‘all-optical’ systems whereby several analogue optical processing modules are cascaded, either via free-space or guided-wave optical paths or interconnections. At the end of the processing cascade the output data rate is presumably reduced to a manageable bandwidth for high-speed electronics. However, one problem with all-optical schemes is the need to store intermediate processor results and to access memory for subsequent processing. Advanced optical memories could be employed to enable all-optical schemes. Section ‘Optical storage’ and chapter 3.4 discuss advanced optical storage concepts such as page-oriented memories and holographic 3D memories, using SLMs and materials such as photorefractives, as described in section ‘Electro-optics’.

C3.5.1.2 Numerical optical processing

To obtain more general applicability of analogue optical processors, various encoding techniques have been investigated to improve the dynamic range and accuracy of such processors, and to overcome limitations of a processor that performs only multiplication and addition. Logarithmic encoding has been employed to compress dynamic range requirements into a smaller analogue voltage range, and to address the difficulty of performing the division operation optically, i.e. by converting division to an addition/subtraction problem. Numerical optical processing for potential implementation of an optical computer has also been considered. Approaches such as residue arithmetic [9] and digital multiplication by analogue convolution (DMAC) [10, 11] have been explored extensively. However, all such approaches involve many nonlinear, logic operations that have to be performed optically to avoid numerous conversions between the optical and electrical domains. Consequently, many all-optical (optical IN, optical OUT) switch devices have been investigated, for performing logic operations, such as nonlinear Fabry–Perot etalons and the self electro-optic effect device (SEED) which is based on the quantum-confined Stark effect [12]. Significantly reduced power consumption in performing these nonlinear operations has been achieved with SEED devices [13] compared to early switches. However, no approach has, to date, resulted in an optical numerical computer competitive with electronic numerical computers. Fundamental arguments have been forwarded about the disadvantages of performing numerical operations optically, such as the large minimum-power requirement to perform nonlinear operations optically compared to electronically, even with SEED devices [14], the limits on integration density in such processors due to lower limits on the size of optical processors due to the wavelength of light [15], and the lower signal-to-noise of optical signals compared to electrical signals because the fundamental shot-noise power limit of optical signals is larger (for the same bandwidth) than the Johnson (or thermal) noise limit for electrical signals.

C3.5.1.3 Optical interconnections/optics in computers

An alternative paradigm for optical implementation of computational algorithms (vice the more familiar arithmetic formulation) is to recast particular signal and image processing algorithms in terms of routing of optical data among various nodes according to a particular interconnection scheme and with various weighting factors for data that recombine at nodes. This paradigm is similar to that used in neural-net formulations of processing problems [16]. The data nodes are ‘neurons’. At these nodes only multiplication and addition are required, the former for applying weighting factors, and the latter at the recombination nodes, respectively. A simple threshold operation at the recombination node(s) residing at the output plane is often required. The advantage of using optics to perform the routing function is based on the free-space propagation characteristics of optical beams. Different optical channels (at reasonable power levels, so nonlinear effects can be ignored) can be routed in free space without interfering. Unlike physical wires, optical channels can overlap within the same space.

A data-routing paradigm can be used to describe the basic Fourier transform operation by a lens or a hologram, as has been described above and illustrated in figure C3.5.1. The Fourier Transform is effected as a global optical ‘interconnection’; the interconnections are global since every input datum (or pixel) is connected to every output (transform) datum, and the input weighting factors are the amplitude and phase values at the pixel locations at front focal plane. A more general formulation involves both global and nonglobal interconnection, such as in neural-network formulations where layers of ‘neurons’, or simple processors that perform addition and multiplication of data, are interconnected. The multiplicative weighting factors applied at various neurons can be adaptively adjusted to solve signal- and image-processing problems [16]. Optics has been explored for implementation of specific neural net algorithms [17, 18]. The dynamic range and accuracy requirement for the additions and multiplications in neural net algorithms is generally not high, and therefore can be performed adequately with analogue data in an optical implementation. However, while optics has potential to implement massively parallel neural networks that might be difficult for electronic implementation, the maturity of neural network theory at this time has not yet progressed to the point where neural processor performance is limited by the size and interconnectivity possible with electronically implemented neural nets. Hence, the intrinsically large optical interconnection capability is not yet needed. Therefore, the alternate development of optical routing and processing capabilities has been towards enhancing the capabilities of computers and telecommunication networks (see chapter C1.2). All-optical digital switching fabrics for interconnection of large fibre-optic networks have been addressed with arrays of SEED devices [19] that provide the logic operations needed for switching. While switching fabrics have been successfully demonstrated [20], the scale of fibre-optic telecommunication switching needs have not yet called for such ultra-high-bandwidth fabrics. Hence, optical interconnection has concentrated on use with within electronic computers, vice optical computing, and towards interconnecting large arrays of sensors and their associated electronic processors.

One avenue of research for using optics in computers has been to explore simply the interconnection among and within all-electronic computing elements in conventional and novel multiprocessor architectures. A second avenue has been to optically perform necessary switching, routing, and pre-processing operations for individual fibre links and networks of computing elements. Many investigations have addressed advanced concepts for interconnection and storage to implement novel multiprocessor and networked computing architectures [21]. Expectations have been driven by the fact that optical communications and optical storage have been the two most successful commercial optical technologies. However, these commercial technologies usually function exactly as in corresponding electronic subsystems. Therefore, significant research is required to adapt commercial technologies to fully exploit the interconnect potential of optics.

Optical interconnection schemes for computers range from computer-to-computer, box-to-box, card-to-card backplane, and finally chip-to-chip within a circuit card. The first of these already exists in several forms commercially; hence, research has concentrated on the latter applications. Guided-wave (also known as lightwave) interconnections are used for high throughput, low power and low bit error rate. The application is generally in novel multiprocessor computer configurations that focus on reducing or eliminating bottlenecks in conventional computers such as relating to memory access time. Free-space optical interconnections can be used advantageously as the number of processors increases, and they allow three-dimensional (3D) multiprocessor configurations [22]. For circuit-card and chip interconnections free-space optical beams do not have the limitations of physical wires, so circuit cards and elements may be laid out with more degrees of freedom.

C3.5.1.4 Interconnection processing/in-fibre processing

In addition to interconnecting network nodes for computer or communication networks, one can consider processing data residing within the interconnect path. Such processing is generally feasible only for fibre or guided-wave interconnections. The data may be either in digital or analogue form, the former for conventional digital computer and telecommunication systems, and the latter for arrays of fibre-optic sensors. While the type of processing will vary depending on the nature of the optical fibre systems, the processing load can be expected to scale with the bandwidth of the data conveyed. With optical modulation rates now in excess of 100 GHz (see section 'Temporal modulation and processing'), commensurate processing speeds are required. Faster processing will allow more flexibility and capability in fibre systems, such as routing and switching a larger number of digital data channels. Use of conventional all-electronic approaches imposes a need to perform optical-to-electronic and electronic-to-optical conversions. Such conversions add complexity, and high-speed electronics will tend to be power hungry and may not provide the throughput required, e.g. due to electronic processor latencies. Hence, it is attractive to perform processing directly on the optical data stream at the transmission rate, e.g. in-fibre processing (see section 'Optical sensors'). The possibility of manipulating the information stream within an optical fibre has been made possible by (1) the development of optical fibre amplifiers (see chapters A1.8 and B5), (2) the capability to build long fibre delay lines needed for short-duration buffer storage and for implementation of tapped delay lines, and (3) fibre couplers for tapping into delay lines and forming devices such as interferometers.

High-bandwidth digital and analogue data on fibre interconnects are often comprised of a large number of lower-bandwidth channels. There is need to multiplex and de-multiplex these channels. Digital systems may eventually require optical implementation of functions now performed electronically; these include clock recovery and reading of packet headers to enable packet switching and routing. However, if optical de-multiplexing is possible, electronic means for clock recovery will usually be available at the lower channel bandwidths. Routing of de-multiplexed channels requires reading of digital packet header information. As has been noted above, it has been difficult for optics to implement a digital computer; thus it is presently difficult to optically perform the logic needed to read packet headers and perform packet switching. However, it appears attractive to optically switch/rout high-bandwidth, fully-multiplexed, light streams, since these can usually be switched at rates much lower than the data bandwidth, and therefore avoids optical-to-electronic conversion at the signal bandwidth (see chapter C1.4). Optical fibre networks (see chapter C1.2) also can possess unique characteristics that require corresponding new processing functions. A major unique aspect is the wavelength-division multiplexing (WDM) capability of most optical networks, which then requires functions such as wavelength selection and amplitude equalization among all the wavelength channels in a WDM system.

Optical processing involving any of the various paradigms described above is practical only if materials and technology exist to construct devices that effectively implement the desired processing operations. The early literature generally showed architecture developments that required device performance exceeding capabilities at the time. Experience has shown that the former must work in conjunction with the development of the latter. Therefore, in the following we describe materials and devices that have been important to the development of optical processing techniques. Specific applications are described immediately after descriptions of device performance.

C3.5.2 Optical devices and processing applications

The needs in optical processing for optical modulation, display, storage and routing are shared with many applications, covered in other chapters in this Handbook (see chapters B5, C2.3, and C3.4). However, optical processors must not only efficiently modulate information onto optical beams, but must also rapidly manipulate the optical data to perform useful information processing. The needs for large temporal bandwidths and high-frame-rate 1D and 2D spatio-temporal optical modulation are clear, and are covered in sections C3.5.2.1, C3.5.2.2 and C3.5.2.4 respectively, below. In addition, wavelength multiplexing can be performed in conjunction with both temporal and spatio-temporal processing, and various wavelengths can be selected or rejected out of either temporal or multidimensional data streams. Various wavelength-processing techniques are discussed in section C3.5.2.3

C3.5.2.1 Temporal modulation and processing

Processing rate can be maximized via either increase in single-channel modulation rate or increase in number of modulated channels. High-speed single-channel modulation can be done either by direct modulation of a semiconductor laser/light emitting diode (LED) or by using modulation devices external to the laser. The latter affords more flexibility to address a range of applications, although it is bulkier than the former. The former is attractive for their simplicity and high electrical-to-optical efficiency. However, modulated LED output power is limited, and there is a theoretical limit of approximately 30 GHz for high-bandwidth modulation with laser diodes, due to relaxation oscillations of charge carrier density in the laser cavity [23].

Temporal modulation with external devices achieve the highest speeds and lowest power in optically-guided lightwave structures such as fibre and channel waveguides (see chapter A2.5). Either phase or intensity modulation devices can be employed.

The basic guided-wave structures for external phase modulation are the channel waveguide and the waveguide splitter where a channel waveguide branches out into two guides in the form of a ‘Y’ [see chapters A1.5 and A2.5]. Two waveguide splitters can be combined to form the guided-wave version Mach–Zehnder interferometer, as shown in figure C3.5.2. Application of a voltage, V, across a channel waveguide fabricated in an electro-optic material such as $LiNbO_3$ will alter the optical path length, or equivalently, the phase of light passing through the waveguide due to the electro-optic effect [24]. If desired, the phase-modulated light can be converted to intensity-modulated light by using the

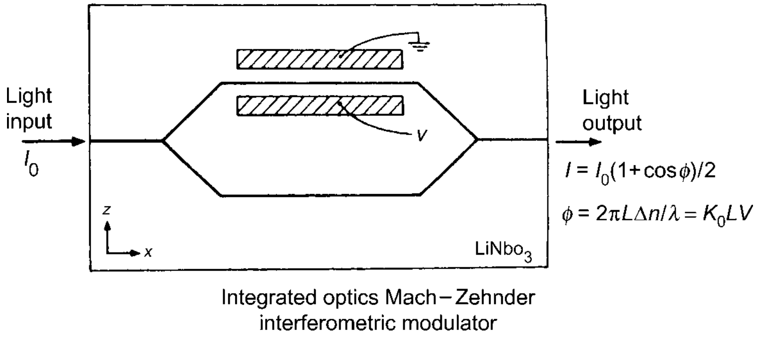

Light
input
I_0
V
z
x
LiNbo$_3$
Light
output
$I = I_0(1+\cos\phi)/2$
$\phi = 2\pi L\Delta n/\lambda = K_0LV$
Integrated optics Mach–Zehnder
interferometric modulator

Figure C3.5.2. Guided-wave Mach–Zehnder interferometer.

guided-wave Mach–Zehnder and, most commonly, modulating the light in one of the two guides. High-speed modulation requires the voltage signal to be applied to a transmission line running parallel to the optical waveguide. The transmission line, which replaces the electrodes in figure C3.5.2, is designed so the instantaneous voltage signal of the transmission line travels at close to the same velocity as the guided optical wave, greatly increasing modulation efficiency. This design approach has produced intensity-modulation devices with speeds and bandwidths of up to 100 GHz at less than 6*V* drive signal [2].

The Mach–Zehnder modulator can be used for analogue intensity modulation. The output intensity as a function of V is a nonlinear but well-known function:

$$I_1/I_0 = 1 \pm \{\sin\pi(V/V_\pi)\}, \tag{C3.5.5}$$

where V_π is the voltage to induce a π phase shift in the light beam, I_0 is the incident light beam intensity and I_1 is the output intensity. The sine-squared function can lead to spurious responses in a broadband signal. The production of spurious third-order signals can be seen from a Taylor-series expansion of the modulated quadrature-point signal

$$1 \pm \sin(\pi V(t)/V_\pi)$$

as

$$1 \pm (\pi V(t)/V_\pi) \pm \{(1/6)(\pi V(t)/V_\pi)^3 \pm \text{higher order terms}\}. \tag{C3.5.6}$$

Hence, if there are two frequency components f_1 and f_2 in the broad band signal, third-order intermodulation signals at $2f_1 - f_2$ and $2f_2 - f_1$ will appear within the band [25]. The effects of the sine-squared transfer function can be minimized by operating two Mach–Zehnder interferometers in parallel or back-to-back [26, 27], but generally at an increase of optical insertion loss.

Fibre-optic versions of a Mach–Zehnder interferometer can be constructed, as shown in figure C3.5.3(*a*). Fused fibre couplers are at both ends of the interferometer legs to split and recombine the light beams from the single-mode fibre. Since silica fibre is not an electro-optic material, an additional component such as a piezoelectric cylinder is used to impose phase shifts onto the light in one leg (figure C3.5.3(*a*)) by increasing/decreasing the length of the fibre. Use of a piezoelectric element maximizes the possible phase change; however, it also limits the modulation speed of such devices.

Intensity modulation can be achieved using electroabsorption in III–V materials such as GaAs and its ternary and quaternary compounds with In, Al, and P [28]. In electroabsorption the optical density of materials whose bandgap is closely matched to a laser wavelength changes with applied voltage. In a device such as the electro-optic Mach–Zehnder device, the modulation voltage changes the optical

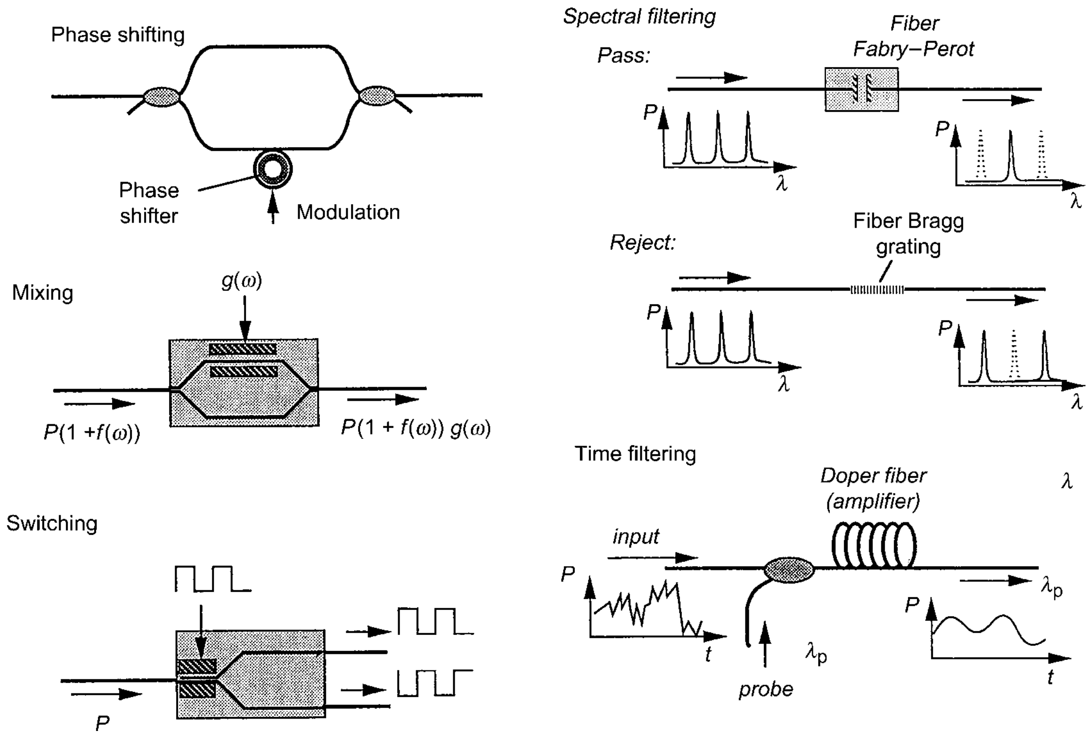


Figure C3.5.3. In-fibre optical processing functions: (*a*) Phase shifting in a Mach–Zehnder device constructed with fused fibre couplers, (*b*) frequency mixing of two signals $f(\omega)$ and $g(\omega)$, (*c*) switching of optical signal between two fibre channels, (*d*) spectral filtering using fibre Fabry–Perot or Bragg grating devices, (*e*) time filtering of a temporal signal using a probe beam at wavelength λ_p.

density (imaginary part of the index of refraction), vice the optical phase (real part of the refractive index) as described above. Use of multiple quantum-well (MQW) materials, with an engineered exciton-line band edge, allows devices with faster operation at lower voltages and with higher signal-to-noise than the classical Franz–Keldysh effect in materials like bulk GaAs. MQW devices consist of a stack of ultra-thin, ~10-nm thick, layers of an alternating composition of III–V materials. Alternating layers of GaAs and AlGaAs are used for modulation at 850 nm wavelength. Alternating layers of quaternary materials such as InAlGaAs/InGaAs are used for operation at 1550 nm. Excitons produced by illumination of the MQW material have a much sharper and narrower exciton-absorption band than in bulk non-quantum-well material. Application of the electric field moves the exciton absorption line towards the red, causing change in the absorption at a fixed laser wavelength. Maximum optical contrast requires operation at the specific wavelength close to that of the bound exciton line in the MQW structure. The contrast change can also be greatly increased by placing the MQW stack within a Fabry–Perot cavity [29], so the incident light makes multiple passes through the stack. For maximum modulation speed a transmission line electrode must be used to maintain synchronism between the modulating signal and the optical signal, as described above for the Mach–Zehnder phase modulator. Since the modulation is due to shifting of a band edge, the modulation is highly nonlinear with respect to drive voltage. Hence, this type of modulator has been primarily used for digital modulation.

It is important to remember that speed, linearity, and efficiency are not the only parameters that need to be considered for temporal modulators. Other parameters that can be important include the capability to handle high optical power, and the device response to variables in the optical beam properties such as polarization state and wavelength. Hence, while the discussion above has concentrated on guided-wave devices, depending on application one may need to consider slower and larger temporal modulators such as acousto-optic devices and micro-mechanical devices such as described in sections ‘Acousto-optic devices’ and ‘One-dimensional electro-optic modulator and applications’ below.

Temporal processing applications in fibre optic systems

Microwave optics and beamforming

Mach–Zehnder waveguide interferometer devices are crucial in the area of microwave optics. The most basic goal in microwave optics is transmission of high-bandwidth analogue signals. Microwave signals are modulated onto an optical carrier for transmission by a fibre-optic line; the advantage of optical transmission is the lower size and weight compared to the use of electrical transmission lines and microwave waveguides. However, care must be taken in order to maximize analogue dynamic range, such as by minimizing third-order intermodulation signals, as discussed above for equations (C3.5.5) and (C3.5.6). Areas of application for microwave-optic transmission include:

- antenna remoting for radar and communications systems, to minimize the weight of transmission lines when antennas must be at some distance from the receiver/transmitter, such as at the top of a mast,
- in wireless networks, for feed lines from base stations to the cellular-network antennas.

In addition to signal transmission, coherent optical devices can be constructed for antenna-array applications such as beamforming. In simplest terms, beamforming is the maximization of the transmit power of an array of antennas (for a transmitting array, or receive power for a receiving array) at particular look angles and frequencies. Minimization of receive sensitivity can also be an objective for rejection of spatially distributed interference. Maximization/minimization of transmit/receive power can be achieved by applying the correct phases to the antenna signals to produce constructive/destructive interference in the desired directions. For microwave-optic devices, the optical phases can be adjusted to produce the desired beams. The attraction of optical beamforming at microwave frequencies is the present difficulty in performing the necessary operations digitally at this high frequency range. To illustrate the beamforming operation we consider the simple case of a 1D linear array of N equally-spaced antenna elements. These elements can be either transmitters or receivers of a microwave signal at frequency f_0.

The phase difference of the signals from/by adjacent array elements at look angle α measured from the perpendicular to the line connecting the array elements is

$$\exp[\mathrm{j}2\pi f_0(d\sin\alpha)/(v)]$$

where d is the array-element spacing, and v is the signal velocity. The phase shift at the nth element relative to the first is therefore

$$\exp[\mathrm{j}2\pi f_0(nd\sin\alpha)/(v)].$$

If the signal strengths at the antenna elements x_n are sampled at a given instant of time and summed, one obtains (with wavelength $\lambda = f_0/v$)

$$b = \sum_{n=0}^{N-1} x_n \exp[\mathrm{j}2\pi(nd\sin\alpha)/(\lambda)]. \tag{C3.5.7}$$

To obtain signal gain, each term must be multiplied by a phase factor. For reasons soon to be apparent, the phase factor is chosen as

$$\exp[-\mathrm{j}2\pi nk/N].$$

The new summation is

$$b_k = \sum_{n=0}^{N-1} x_n \exp[\mathrm{j}2\pi(nd\sin\alpha)/(\lambda)]\exp[-\mathrm{j}2\pi nk/N] \tag{C3.5.8}$$

which is maximized when the exponential factors in every term cancel to produce unity, i.e. when

$$k = (Nd\sin\alpha)/v.$$

The quantity k is thus a phase corresponding to a beam at look angle α. The choice of phase factor for producing the beam b_k is be seen to be equivalent to a spatial discrete Fourier transform (DFT) operation on the signal samples received from/by the transmit/receive array, respectively.

The above beamforming example applies best to a continuous-wave signal with small fractional bandwidth centred at microwave frequency f_0. If the microwave signal is short, e.g. an impulse function, as commonly for radar, the above beamforming algorithm is not optimum, since the signal has large fractional bandwidth $\Delta f/f_0$. In this large bandwidth case, time delays rather than phase shifts need to be applied to the array elements to produce a beam $B(t)$.

$$B(t) = \sum_{n=0}^{N-1} a_n x_n(t-\tau_n) \tag{C3.5.9}$$

where τ_n is the delay applied to the nth element in an array of size N, and a_n is a weighting function for shaping the angular characteristics of the beam, e.g. half-intensity mainlobe width and sidelobe levels [30]. The required time delays between adjacent elements is given by

$$\tau_{n+1} - \tau_n = (d\sin\alpha)/v. \tag{C3.5.10}$$

A fibre delay line for each array element can be used to implement the required delays, but this approach lacks the flexibility to easily change the beam direction. A novel alternate approach is to use different optical wavelengths for signal transmission to each array element. The wavelength-dispersion characteristics of the optical fibre, where different wavelengths will have different time delays over the same length of fibre [31], then contribute to the time delays between adjacent array elements.

The 1D example generalizes to unequal spacings between array elements and to 2D arrays in a straightforward manner.

Optical sensors

The Mach–Zehnder device, either fibre or waveguide version, is also an important class of device for sensors (see chapter C3.1). For example, one leg of a fibre Mach–Zehnder is coated with a material that will change the length of the fibre in response to a specific external stimulus such as stress/strain, temperature, magnetic field, etc. Both the waveguide and fibre-optic Mach–Zehnder interferometers can perform a number of temporal operations on the optical signal: phase shifting (figure C3.5.3(*a*)), mixing of a modulated signal with another (figure C3.5.3(*b*) — note that the modulated input signal must be applied with a bias), switching of the polarity of signals (figure C3.5.3(*c*)), spectral filtering using either a fibre Fabry–Perot or a Bragg grating (figure C3.5.3(*d*)), and sampling or time filtering of high-bandwidth signals (figure C3.5.3(*e*)). Note the necessity for optical amplification, and its implementation in the optical domain with fibre amplifiers, as shown in figure C3.5.3(*e*), where the fibre is doped with material such as erbium and the probe beam can also activate the amplification. When a similar fibre-optic Mach–Zehnder is used for demultiplexing and demodulation of sensor array data, higher accuracy is achieved because of compensation for the nonlinear (1 + sine) transfer characteristic of a single Mach–Zehnder device, as already discussed above for equations (C3.5.5) and (C3.5.6).

Optical interconnect for computers

Parallel high bandwidth optical channels, either guided-wave or fibre-optic, have been explored for optical interconnection applications [32], thereby multiplying the throughput rates that can be achieved on a single channel. However, simple aggregation of a multiplicity of single-channel optical-link hardware quickly becomes impractical, especially in applications where volume and power must be limited, e.g. within a computer. Fibre-optic 'ribbon' cabling, consisting of a number of closely spaced fibres, is thus usually chosen to interconnect a number of high-speed processors or boards. Challenges in producing such optical interconnect hardware include the monolithic integration of arrays of lasers, modulators, and detectors. Achieving the requisite density for laser arrays that can be coupled to fibres has necessitated the development of vertical-cavity surface-emitting laser arrays (VCSELs) with low power threshold for lasing and high electrical-to-optical efficiency [33]. The required laser efficiency has been achieved through use of quantum well materials that reduce the number of allowed excited states that must be optically pumped [34].

Wavelength processing in fibre optic systems

WDM technology has been a major development in optical fibre telecommunications systems. Figure C3.5.3(*d*) illustrates how wavelength channels can be passed or rejected using, respectively, fibre devices such as the fibre Fabry–Perot filter, a mechanical device (chapter 8.5), and fibre Bragg gratings (chapter xxxx), refractive index gratings written perpendicular to the long dimension of the fibre via the photorefractive effect. While WDM technology is primarily used in long-haul telecommunications to increase capacity without commensurate increase in physical plant, WDM technology can also be applied to short distance optical interconnects to provide an additional degree of freedom. For example, in ribbon fibre interconnects, each fibre can have a separate wavelength. The wavelength of a channel can be used as an identifier for data routing, reducing the need, or even avoiding the need to read the packet header (a difficulty for optics mentioned in section C3.5.1.4). A key enabler of WDM optical interconnects is the monolithic multiple wavelength surface emitting laser array [33], which provides either a unique wavelength for each laser, or fewer but redundant wavelengths.

Free-space optical application

Nonguided-wave temporal modulators can be applied to free-space communications, distributed computing architectures, laser radar, and laser designators. In the latter the use of high bandwidth coding is often needed for increased detection margin or range, using correlation methods, mathematically described in section 'Acousto-optic applications'. For free-space optical communications bulk electro-optic modulators often require high drive powers and have low contrast. However, a novel quantum-well modulator device holds promise for both high speed and high contrast [35]. The modulator is positioned on the entry/exit facet of a corner cube retro-reflector. Electrical signals applied to the retro-reflector result in modulation of an incident laser beam. Data rates of up to 10 Mbps over several metres at bit error rates of 10^{-6} have been demonstrated; the approach has the potential for hundreds of megabits/second at power consumptions below 100 mW. Use of two retro-reflectors allows a two-way link to be established.

C3.5.2.2 One-dimensional spatial light modulators and applications

Acousto-optic devices

Optical modulation of data in a 1D spatial format allows parallel processing of blocks of temporal data. The most effective 1D light modulators have been acousto-optic, and this technology has played an important role in the demonstration of optical processing architectures. The basic construction of an acousto-optic device is shown in figure C3.5.4. A high-bandwidth rf drive signal is applied to an acoustic, piezoelectric, transducer that has been bonded to one end of the acousto-optic cell, using acoustic impedance-matching materials. The resultant acoustic wave is a replica of the rf drive signal. The rarefactions and compressions of the acoustic wave produce corresponding refractive-index changes due to the elasto-optic effect. The cell thus contains a travelling phase grating corresponding to the acoustic wave. The main features of acousto-optic diffraction important to optical processing are now summarized.

- The maximum diffraction efficiency, into a single order, occurs when the difference between the momentum vectors of the incident and diffracted light is equal to the acoustic-wave momentum vector. For an isotropic material momentum matching occurs when light is incident to the acoustic wavefront at the Bragg angle θ_B, defined by

$$\sin\theta_B = \lambda/(n\Lambda) = \lambda f(2nv) \qquad \text{(C3.5.11)}$$

  where Λ, f, and v are the acoustic-wave wavelength, frequency and velocity, respectively, λ is the optical wavelength of the monochromatic incident light, and n is the index of refraction of the medium at λ. To diffract light predominantly into only a single diffraction order one must examine the quantity Q,

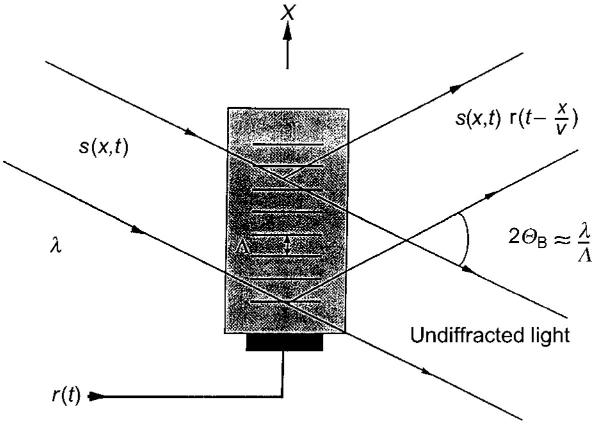


Figure C3.5.4. Construction of acousto-optic cell and Bragg cell geometry for data multiplication.

$$Q = \frac{2\pi\lambda Z}{n\Lambda^2}.$$

For $Q > 7$, analysis shows that the first diffraction order contains $>90\%$ of the diffracted light [36].

- The diffraction efficiency is given by a nonlinear sine-squared relationship, that arises from considering the diffracted (I_1) and undiffracted (I_0) light as two modes of a coupled mode system [37].

  $$I_1/I_0 = C_0 \sin^2\{M_2(\pi^2/2\lambda^2)(L/H)C_{\mathrm{rf}}P_{\mathrm{rf}}\}^{1/2} \qquad \text{(C3.5.12)}$$

  where P_{rf} is the rf drive power, L/H is the ratio of width-to-height for the acoustic transducer, C_0 and C_{rf} are constants, and M_2 is a figure-of-merit for diffraction efficiency that depends only on material parameters, and is given by

  $$M_2 = (n^6 p^2)/(\rho v^3) \qquad \text{(C3.5.13)}$$

  where n is the refractive index, p is the elasto-optic coefficient, ρ is the density and v is the acoustic velocity of the material. The sine-squared relationship, equation (C3.5.12), leads to similar considerations as for the Mach–Zehnder devices discussed earlier, equation (C3.5.5), such as limitation on wideband dynamic range due to third-order intermodulation products. Also, I_1/I_0 must be less than 3% in order for the relationship between I_1 and P_{rf} to be linear to better than 1%.

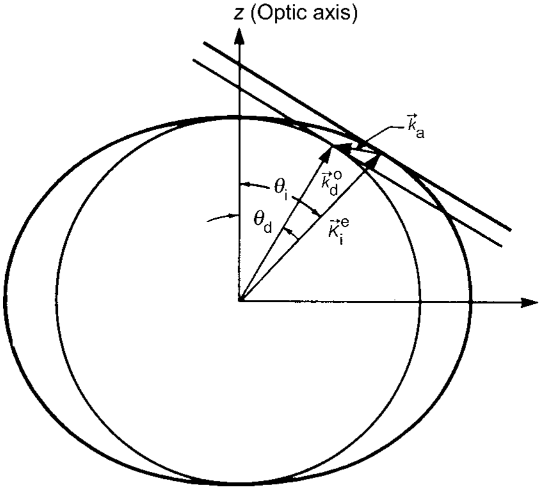
Extensive literature exists on materials and design of acousto-optic devices [38, 39]. The figure-of merit M_2, equation (C3.5.12), provides guidance on choice of optimum AO device material, especially dependence on refractive index n. Device development has converged on $LiNbO_3$ for operation at high frequencies and modest time-bandwidth (up to 4 GHz, ~100 TBW), TeO_2 for lower frequency application but large time-bandwidth (<100 MHz, several thousand TBW), and GaP for intermediate bandwidths and time-bandwidths (up to 1 GHz, several hundred TBW) [6].

The major aspects of acousto-optic devices are summarized below:

- For the isotropic case, the diffracted light is deflected by an angle $2\theta_{\mathrm{B}}$. The more general nonisotropic diffraction case is shown in figure C3.5.5, compared with the isotropic case. In this nonisotropic case, the incident light is polarized as an extraordinary ray with momentum vector $\boldsymbol{K}_{\mathrm{i}}^{\mathrm{e}}$, and the diffracted light is of opposite polarization from that of the incident light, $\boldsymbol{K}_{\mathrm{d}}^{0}$. The angular relationship among the incident and diffracted light beams and the acoustic vector $\boldsymbol{k}_{\mathrm{a}}$ is determined by the necessity for each beam to reside on its respective optical-index ellipsoid, and is shown in figure C3.5.5.

- The information in the cell is constantly updated due to the travelling wave nature of the acoustic signal. If the cell is driven by an electrical signal $r(t)$, the information displayed along the cell direction, x, is given by $r(t - x/v)$, as shown in figure C3.5.4.

- Complex (both amplitude and phase) data on an incident light beam, $s(x,t)$, that spatially fills the cell aperture will be diffracted by the acoustic grating and therefore contain the product of $s(x,t)$ and the complex data on the acoustic wave, $r(t - x/v)$. The amount of spatial data contained on the modulated beam is equivalent to the number of resolvable deflection position for the optical beam and is the so-called time-bandwidth (TBW) product, equal to the product of the length of the cell, in units of time, and the temporal bandwidth of the drive signal [8]. This result is easily derivable from equation (C3.5.11) and the angular optical-diffraction limit due to the finite spatial extent of the incident light beam.



Wave-vector diagram for noncollinear acousto-optic filter.

Figure C3.5.5. Anisotropic Bragg diffraction geometry.

- If the continual temporal update of data in the cell, represented by $r(t - x/v)$, is not desired, a short pulse of light can be used to 'freeze frame' the instantaneous contents of the cell. Alternatively, the incident light beam can be focused into a small diameter spot within the cell, resulting in a temporal modulator [40]; the bandwidth of such a temporal modulator is determined by the traversal time of the acoustic wave through the spot of light, and initiation of modulation is determined by the delay caused by transit of the acoustic wave from the transducer to the optical spot location.
- The centre frequency of the diffracted light is shifted relative to the incident light frequency by the carrier frequency of the acoustic signal. The frequency shift arises from energy and momentum conservation considerations in diffraction by a moving grating. Either frequency upshift or downshift is possible, depending on the momentum direction of the incident light relative to the acoustic beam momentum direction. Acousto-optic diffraction is an important means of producing small frequency shifts onto the ~100 THz optical carrier.

Multi-dimensional devices using acousto-optics

To better exploit the two spatial dimensions of optics, acousto-optic devices have been extended to 2D architectures either by, constructing an array of 1D devices in a single crystal, or by constructing a device with acoustic transducers on orthogonal edges of a crystal. In either of these approaches a large, high quality crystal is required. For multichannel acousto-optic devices, a large number of channels is desirable, but in the absence of electrical crosstalk, the major fundamental limitation is acoustic diffraction. Acoustic diffraction depends not only on the dimensions of the transducer, but also on material. Ideally, the near-field acoustic wavefront is planar over the aperture of the transducer, and diffraction effects are observed in the far-field. Hence, it is desirable to maintain a near-field condition over as long a propagation distance as possible. In an isotropic medium, the transition from near to far field occurs at a distance from the transducer of approximately

$$D = H^2/8L \tag{C3.5.14}$$

where H is the vertical dimension of the transducer and L is the acoustic wavelength. However, for an anisotropic medium, D can be increased by an additional factor of $c = (1 - 2b)^{-1}$. The quantity b is the coefficient of the q^2 term in a power series representation of the acoustic slowness surface [41],

$$K_a(q_a) = K_a(1 + bq_a^2 + dq_a^4 + \cdots) \tag{C3.5.15}$$

where q_a is the acoustic-wave direction relative to the normal to the transducer. For shear mode TeO_2 the quantity c is 0.02 so that the acoustic spreading is very rapid. Because of the severe acoustic spreading in shear mode TeO_2, multichannel devices have been constructed only of longitudinal mode TeO_2 where the spreading factor c is 2. A 32-channel TeO_2 device at a centre frequency of 250 MHz has been demonstrated [42, 43].

The second 2D approach uses orthogonally-propagating acoustic waves. Using two separate 1D cells orthogonal to each other, a set of anamorphic optics passes light from one cell to the other, where the data in the first cell are focused and passes through every point on the second cell. This scheme is illustrated in figure C3.5.6. A much more compact alternative to individual 1D devices is to employ a single large square crystal with transducers along two edge facets and light propagating perpendicular to the square aperture. However, this approach is feasible only in cubic materials. TeO_2 satisfies this criterion, and shear-mode TeO_2 devices have been demonstrated. However, if the two acoustic waves propagate within the same volume of crystal, nonlinear mixing of the signals will occur at lower power levels; this is particularly true for TeO_2, which exhibits acoustic nonlinearities at relatively low power levels. Thus, devices are made with the acoustic transducers offset from each other along the optical-beam direction. The transducer dimension along the optical path and the offset distance must be minimized to prevent loss of resolution or to avoid use of optics with large depth of focus.

Acousto-optic applications

Analogue processors

One-dimensional processing. The most basic application of the acousto-optic cell is to place information at the front focal plane of a Fourier Transform lens. The spatial spectrum of the travelling wave is then displayed at the back focal plane. Since acousto-optic devices have been constructed into the microwave frequency range, up to 4 GHz [6], an rf spectrum analyser may be constructed as shown in figure C3.5.7. A collimated optical beam with intensity profile $a(x)$ illuminates an acousto-optic cell P1 that is driven by the signal $s(t)$. The signal $s(t)$ must first be mixed with a carrier signal $\cos(2\pi f_c t)$ to produce the drive

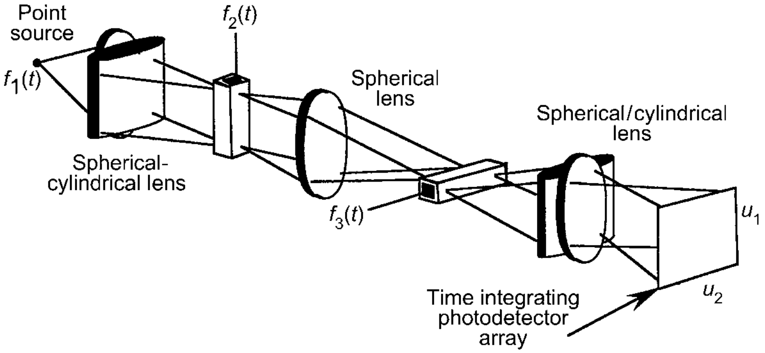

Point source
$f_1(t)$
$f_2(t)$
Spherical lens
Spherical/cylindrical lens
Spherical-cylindrical lens
$f_3(t)$
u_1
u_2
Time integrating photodetector array

Figure C3.5.6. Triple-product processor using orthogonal acousto-optic Bragg cells.

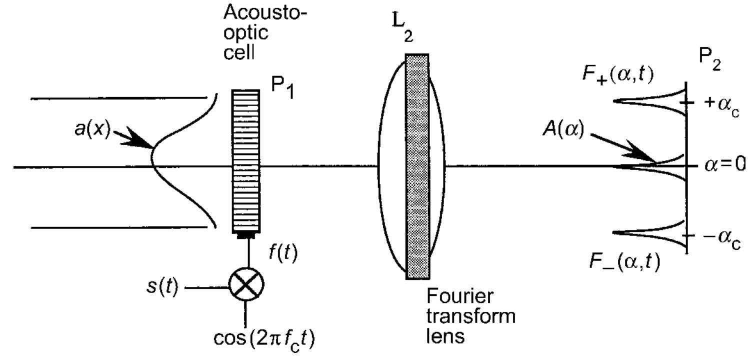


Figure C3.5.7. Acousto-optical spectrum analyser.

signal $f(t)$ at the acoustic frequency f_c. The complex spectrum of $f(t)$ appears at the back focal plane P2 of the Fourier transform lens. The spectral components with frequency upshift are represented by $F_+(\alpha,t)$, and with frequency downshift by $F_-(\alpha,t)$. The Bragg diffraction geometry determines whether F_+ or F_- is displayed. The number of resolvable spectral positions is equal to the TBW product, as described above. The spectral update rate is approximately equal to the length of the illuminated portion of the Bragg cell, typically in the sub-microsecond to ten-microsecond range. The photodetector array at the back focal plane must therefore have corresponding output frame rate. The spectral information must also be measured with high fidelity and dynamic range, placing additional demands on the dynamic range performance of the photodetector array [44]. For a power spectrum analyser only the amplitude $A(\alpha)$ is measured by square-law photodetector elements. If the phase information must be preserved, each photodetector element must be illuminated with a reference optical beam and the bandwidth of each element must support the difference frequency between signal and reference beam.

The back focal plane information of a spectrum analyser can be cascaded into a second Fourier Transform lens, producing an optical image of the spatial information. Such an optical image can be multiplied with another information array by locating a second acousto-optic cell at the image plane, as shown in figure C3.5.8. The travelling-wave nature of the acoustic-cell data has been advantageously used to implement the correlation integral, equation (C3.5.4). A final lens performs the Fourier Transform of the product of the functions shown in figure C3.5.8, and the output of a small detector at the centre of the back focal plane is the correlation integral

$$R(t) = \int f(t + x/v)s(t - x/v)\mathrm{d}x = \int f(u)s(u - 2t)\mathrm{d}u. \tag{C3.5.16}$$

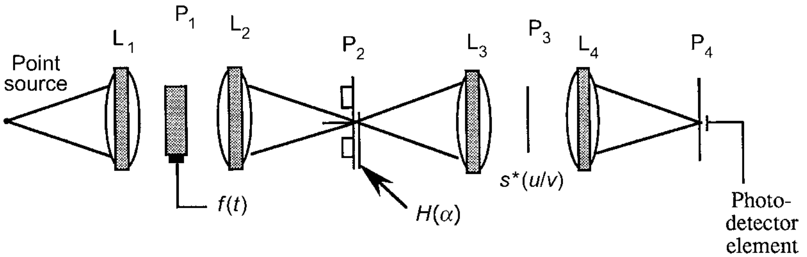


Figure C3.5.8. Space-integrating acousto-optical correlator.

The arrangement of figure C3.5.8 has been used to perform high-speed correlation according to equation (C3.5.16) [7]. The integration can be thought of as performed by the focusing action of the lens, hence the correlation architecture is termed ‘space-integrating’. Processing of a long-duration signal with the space-integrating architecture requires formatting of input data into frames of size, or length (when inserted into the cell), less than that of the illuminated portions of the cells. This formatting adds to the burden for high speed pre-processing in this processor.

An alternative to spatial integration using a lens, as shown in figures C3.5.7 and C3.5.8, is to employ time-integration in the photodetection process [45]. The product terms within the kernel of integral transforms such as equations (C3.5.4) and (C3.5.16) can be obtained by successive diffraction by acousto-optic cells of modulated light. The duration of signals is not limited by the size of the cells and photodetector array, but by the feasible integration time. The photodetector array does not need to be read until its charge capacity is reached, and digitization of the output allows further signal integration time. Output data rates in these ‘time-integrating’ architectures are much lower than with the space-integrating architectures. However, only a small fraction of the resultant correlation function is produced, and some a priori knowledge of the location of the region of interest is required. Otherwise, the amount of output data does not decrease.

Multi-dimensional processing. A multiplicity of 1D spectrum analyser and correlation processors can be implemented in parallel using the multichannel acousto-optic devices described in section ‘Multi-dimensional devices using acousto-optics’. The main application of these multi-channel processors is to process data from arrays of sensors. Array functions are performed in addition to the 1D analysis; these functions include:

- Direction-finding in conjunction with spectral analysis [44] using an array of rf receive antennas. Each array element drives a separate channel of a multi-channel acousto-optic device where the acoustic transducers are relatively spaced exactly as for the array antennas. By coherently illuminating the multiplicity of channels, the phase differences among the various channels result in additional deflection orthogonal to the spectrum-analysis direction that provides information on the angle-of-arrival of the signal at the antenna array.
- Generalized multi-dimensional array beamforming [46]. The acousto-optical channels impose phase factors or time delays on array signals to produce directional transmit or receive beams, using a straightforward extension of the formalism described in section ‘Temporal processing applications in fibre optic systems’ for a 1D array.

A more general extension of Fourier transform processing into two dimensions using acousto-optic input devices has been explored. With the orthogonal-cell arrangement shown in figure C3.5.6, one can perform (1) 2D beam deflection in free space, for applications such as optical interconnection and raster addressing in displays, and (2) 2D processing where the processing kernel consists of two 1D factors. For the latter, the two orthogonal acousto-optic devices are combined with a temporally-modulated light source and a 2D integrating photodetector array to produce what is known as a triple-product processor (TPP) [47]. As illustrated in figure C3.5.6, the time-integrating version of the TPP output has the form

$$g(u_1, u_2) = \int_0^{kT} f_1(t) f_2(t + u_1) f_3(t + u_2)\,\mathrm{d}t \tag{C3.5.17}$$

where u_1 and u_2 are the coordinates of the output plane, T is the maximum integration time of the product signal onto the 2D photodetector array, and k is a constant ≤ 1 depending on the duration of the input signals. By employing various functions f_1, f_2, and f_3, the TPP, equation (C3.5.17), can implement a number of important signal-processing operations at extremely high speed:

- Time-frequency transformation, such as Ambiguity processing (also known as range-Doppler processing) in radar and sonar [48]. The product of the functions $f_1(t)$ and $f_2(t+u_1)$ constitute the cross-correlation kernel, as in equation (C3.5.4), to provide range determination. Function f_3 is the Fourier phase factor for compensating for any Doppler effects on the cross-correlation function, thereby resulting in the 2D ambiguity surface. Additional 2D time-frequency transforms that can be performed with the TPP include the Wigner function, instantaneous power spectrum, and cyclostationary function [46].
- Fourier transformation of very large 1D blocks of data in 2D format, also known as folded spectrum analysis [47]. Since each cell of the TPP can have TBW of about 1000, a signal can be spectrally analysed into approximately one million narrow frequency bins.
- Matrix operations [49]. Matrix operations are employed, not for numerical processing, but to allow use of linear algebraic techniques to formulate an optical implementation of various signal-processing operations for the TPP. For example, analogue multiplication of three matrices has been implemented by cascading the TPP into a 2D SLM (covered in section C3.5.2.4). In particular, the triple-matrix product allows the implementation of similarity transforms to be performed; these transforms, which diagonalize an input matrix, perform the important data-reduction function that permits lower-bandwidth results to be passed to subsequent processing.
- Synthetic aperture radar (SAR) image formation [50]. SAR image formation is historically important as one of the first successful applications of optical signal processing. The original processors as input film onto which the radar return signals were written. For near-real time image formation the film can be replaced by a 2D SLM (discussed in section C3.5.2.4 and chapter C.2.3). However, the use of acousto-optic cells allows more-rapid generation of SAR imagery over either film or present 2D SLMs. The original method where the radar return signals were written onto film and the developed film used as the input into an optical system. The ability to use 1D acousto-optic cells to produce 2D SAR imagery follows from recognition that the SAR image-formation equation is separable into factors involving integration over only the range variable η, followed by integration over only the azimuth variable ξ. The equation governing the formation of an image point at coordinates (x,y) is given by

  $$g(x,y) = \int_0^{A1}\int_0^{A2}\left\{f(\eta,\xi)\exp[\mathrm{j}(\pi/\lambda D_1)(\eta+y)^2]\mathrm{d}\eta\right\}\exp[\mathrm{j}(\pi/\lambda D_2(y))(\xi+x)^2]\mathrm{d}\xi \qquad \text{(C3.5.18)}$$

  where $f(\eta,\xi)$ represents the signal from point reflector on the ground, and D_1 and $D_2(y)$ are parameters or functions related to signal curvature (SAR systems transmit a train of short impulses; the resultant radar returns are chirp, or quadratic phase functions, but the chirp rates are different for the η and ξ directions). In the limits of integration $A1$ is proportional to the integrated number of radar pulses in the azimuth direction, and $A2$ determines the duration of the radar pulse train in the range dimension.

The analogue processors described in this section have found only niche application to date. However, a number of prototype acousto-optical processors have exhibited some advantageous features. The advantages generally lie in the size, weight and power consumption advantages over a comparable all-electronic implementation of equivalent processing power. For example, analogue data multiplication using acousto-optic technology at eight-bit precision can have a 350X power consumption advantage over digital multipliers [51]. Specific examples of compact processors include a 6 in^3 rf spectrum analyser [6], and a SAR image-formation processor in $< 1\,\mathrm{ft}^3$ [52]. But the need for significant high speed pre- and post-processing electronics (as described in sections C3.5.1 and C3.5.1.1) results in significant disincentive to use optical processing unless the size, weight and power advantages are paramount. Without such advantages even the early use of optics for SAR image formation has been largely replaced with digital electronic processing. An added measure of advantage could be obtained for optical processors if many of the electronic functions can be performed optically or if, at the output of an optical processor, the amount of data for subsequent electronic processing can be reduced. If a data-reduction algorithm can be implemented, such as the similarity transform mentioned above, one would not need to read out an entire 2D array or could use a linear array; however, this situation is clearly specific only to a special class of data processing. An alternative is to construct a 'smart' photodetector array that selects only the region of interest for readout, such as a local peak [15]. Increasing the variety of operations that can be performed opto-electronically would also allow intelligent algorithms to be implemented. Many all-electronic approaches can be intelligently formulated to reduce the computation load required for a given processing task; use of the fast Fourier transform (FFT) algorithm to greatly reduce computation in digital electronic processing is an example that cannot yet be duplicated in optics.

One-dimensional electro-optic modulator and applications

An electro-optic modulator is a potential alternative to the acousto-optic device. A phase diffraction grating is produced by the electro-optic effect rather than a pressure wave. Such modulators constructed of $\mathrm{LiNbO_3}$ have been demonstrated for high-speed printing applications [53, 54]. However, in the most common configuration the electro-optic modulator requires many parallel addressing lines. Hence, there is significant difficulty in generating a multiplicity of phase gratings. Further, significant effort is required to format data for parallel addressing. However, careful design of the electrical addressing (the intrinsic speed of the electro-optic effect can be sub-picosecond) can result in a device that creates a grating much more rapidly than acousto-optic devices. An alternative approach to parallel electrodes is to use a high-speed shift register to address the electro-optic material. This approach is architecturally similar to that of acousto-optics. Use of a shift register has advantages such as capability to vary clocking rate.

One-dimensional mechanical modulators and application

Spatial modulation can be performed using micro-electromechanical structures (MEMS). Devices are constructed entirely of silicon for both the addressing circuitry and the modulation structure. The micromechanical modulator design consist of support structures and cantilevered beams [69] or torsion beams extending from the structures; these are all fabricated out of silicon by an etching process [70]. A silicon circuit below the micromechanical structure provides addressing and activation voltages that deflect the structure. The deflection changes the angular position of the reflected light beam, producing intensity modulation in the readout system. The early devices were called deformable membrane devices (DMDs), and this terminology persists although no longer limited to the original membrane MEMS approach.

The importance of MEMS devices has been in their application to all-optical switching for fibre-optic telecommunication systems, where switch speed is not important. The advantage is that the data remain in optical form during the switching operation.

C3.5.2.3 Wavelength processing

The bandwidth of optics can be exploited more fully by using a broad band of the optical spectrum and by subdividing an optical signal into many narrow-wavelength data channels and processing each channel individually. Processing in the wavelength dimension has already been introduced in section ‘Free-space optical application’ with respect to the importance of WDM fibre-optic systems. However, processing in the wavelength dimension can be applied to both temporal and multi-dimensional data. Wavelength-selective acousto-optic and electro-optic devices can be applied to spectroscopic systems, to channelize temporal data, and to perform 1D, 2D and 3D imagery analysis and display.

Acousto-optics

The separation of a broadband optical signal into narrow wavelength channels can be performed using acousto-optics. Many optical modulators are sensitive to the wavelength of operation. However, acousto-optic modulators are unique in that, without a large degree of modification, they can be used as flexible elements that can select a number of wavelengths simultaneously. The phase gratings produced by the acoustic wave will, as any grating, select out specific wavelengths at the appropriate angle of incidence for Bragg interaction. The grating period, corresponding to specific wavelength to be selected, is generated by driving the acousto-optic device with the appropriate rf frequency according to equation (C3.5.6). This spectral filtering action is notable in that by driving the device with a composite rf signal a multiplicity of wavelengths may be selected simultaneously, unlike optical glass and interference filters and resonant tunable structures such as Fabry–Perot cavities.

The resolving power of a grating,

$$R = \Delta\lambda/\lambda$$

is equal to the number of grating lines. Hence, for an acousto-optic tunable filter (AOTF), depending on whether the incident optical beam is parallel or perpendicular to the grating lines, the resolving power is then

$$R = D/\Lambda \text{ or } R = L/\Lambda, \text{ respectively,} \tag{C3.5.19}$$

where D is the extent of the illuminated region (parallel case, beam transverse to the acoustic wave) and L is the length over which the optical and acoustic wave overlap (perpendicular case, beam collinear with the acoustic wave).

AOTFs generally operate in the anisotropic diffraction mode, as shown in figure C3.5.5, so that the filtered output light is orthogonally polarized to the input, i.e. the input beam must also be polarized to allow high contrast isolation between input and output optical beams by using crossed polarization. Good AOTF design for high spectral resolution must maximize

$$R = L(\Delta n)/\lambda$$

where Δn is the maximum difference in refractive indices for the birefringent material [71]. The rf frequencies required for specific wavelengths is governed by phase matching conditions for anisotropic diffraction [41, 55]. Popular materials for AOTFs are $LiNbO_3$ and TeO_2 for the visible to mid-wave IR, Tl_3AsSe_3 for near- to long-wave IR, and quartz (SiO_2) and MgF for the UV. Both bulk and integrated-optic devices are possible. The latter are naturally compatible with optical fibre, and utilize the polarization anisotropy in thin-film optical waveguides in addition to material anisotropy [56].

The major opportunity for AOTF in information-processing applications is in their potential employment in WDM techniques and systems. For conventional fibre optic systems guided-wave AOTFs have been developed to filter the various WDM channels at moderate, microsecond, speeds [56]. However, significant interest has been generated in the use of AOTFs to flatten power across the WDM spectral band in long-distance communications networks. Since the erbium-doped fibre amplifiers used in long-distance networks do not amplify uniformly across the WDM band, and networks need to be dynamically reconfigurable, a rapid means is required to pre-condition the channel powers such that all channel power levels are always equal following the fibre amplifier, allowing proper functioning of the WDM network [57]. In the future, just as in fibre-optic communications, developments in optical interconnections are expected to encompass WDM techniques. AOTFs presently represent a viable wavelength switch for initial development.

The multiple-wavelength capability of AOTFs allows numerous other applications including:

- Color generation and correction in multi-dimensional laser displays [58], such as for laser light shows. 3D displays systems have been developed using the rapid wavelength tunability and scanning capability of acousto-optic devices in conjunction with commensurate optical-beam scanning within a 3D medium such as with a rapidly spinning turbine blade.
- Scene imaging in many spectral bands, i.e. creation of image ‘cubes’ for multi- and hyperspectral techniques for remote sensing in commercial, astronomical, and military applications.
- Wavelength-selective microscopic probing for research in biological processes, such as by selectively quenching flouresecence in biological samples with specific wavelengths [59].

Some of the present development issues in the use of AOTFs include, materials and devices for operation further into the UV and IR bands, power requirements due to the necessity for the rf drive, spectral sidelobe levels which limit channel selectivity in WDM networks, and imaging quality through AOTF devices.Finally, the Bragg diffraction requirement often limits the angular acceptance angle for imaging, and the use of rectilinear shapes for the AOTF can introduce astigmatism into the imaging system.

Electro-optics

Electro-optic devices can produce phase gratings analogous to those in acousto-optic devices, as previously mentioned in section ‘One-dimensional electro-optic modulator and applications’. Electro-optic devices have also been considered for wavelength filtering. By laying an array of electrodes onto a guided-wave structure, optical wavelength filtering can be performed analogous to the integrated-optic. However, it is desirable to have devices that are faster than either the electro-optic and acousto-optic devices discussed so far. One approach is to employ a bulk device with volume holograms as the wavelength-sensitive element.

Volume diffraction gratings can be produced optically in a number of materials using the photorefractive effect [60, 61]. The photorefractive effect is a manifestation of the electro-optic effect on a microscopic scale. Internal space-charge electric fields, E_{sc}, arise from photocharge generation, followed by separation of the charges via photoconduction (transport of charge from illuminated to unilluminated regions) or via application of an external field, E_{dc}. Gratings are produced from an interference pattern obtained with two intersecting coherent optical beams. However, a rapid means is needed for either changing the grating period or changing to a different grating. Presently, sub-microsecond photorefractive processes are possible in quantum-well materials [3], but packet switching requires even faster speeds.

A recent development, entitled electroholography has potential to achieve nanosecond switching [62]. The quadratic electro-optic effect is used in paraelectric crystals such as $KTa_xNb_{1-x}O_3$ (KTN). In addition to a space-charge field E_{sc}, a dc field, E_{dc}, is applied, leading to a cross term $E_{sc}\,E_{dc}$. Until E_{dc} is applied, the information contained in the hologram is latent, the diffracted light containing only bias (E_{dc}^2) and unipolar (E_{sc}^2) terms. In principle, such switchable gratings would allow processing operations required in packet switches (e.g. multicasting, power management, data monitoring) to be performed at close to the data bit rate, entirely in the optical domain.

C3.5.2.4 Two-dimensional spatial light modulators

2D SLMs are the most critical devices required for full and efficient exploitation of the inherent parallel-processing and interconnection capabilities of optics. As discussed above, and shown in figure C3.5.6, 1D acousto-optic devices have successfully performed a number of important 2D signal-processing operations. However, that approach will work only for problems that can be separated into two 1D factors. Many image processing operations cannot be so factored.

Although chapter C2.3 discusses SLM devices and technology, this section reviews necessary and desired performance characteristics for information processing. Devices will be cited without repeating descriptions already provided in chapter C2.3. We first briefly review the major types of SLM.

Two-dimensional SLM devices categories

The first level of categorization is whether the device is electrically-addresssed (E-SLM) or optically-addressed (O-SLM). The next level of categorization is by modulation mechanism or material. Within each category both O-SLMs and E-SLMs are possible.

- Electro-optic. These use traditional linear electro-optic materials such as $LiNbO_3$, KD*P, and bismuth silicon oxide, and quadratic electro-optic material such as PLZT. There is usually a strong trade-off between frame rate and resolution.
- Liquid crystal. There are two major classes.
    - Twisted nematics, as used in commercial displays. Large arrays are available ($\sim 10^7$ pixels), but frame rate is limited by the natural relaxation time of the liquid crystal, of the order of 10 ms.
    - Ferroelectric liquid crystals (FLCs) possess a permanent electric dipole. Both write and erase speed can be increased in proportion to drive signal level. Frame rate can therefore be high, limited by addressing circuitry. The most mature devices are binary contrast only. Analogue devices with large grey-scale capability are less mature
- MQW structures. These devices employ the quantum confined Stark effect in stacks of ultra-thin, $\sim$10-nm thick, layers of III–V materials, generally of an alternating composition. Quantum well materials have intrinsically very fast response times, and have been described in section C3.5.2.1 and in chapter B14. Large 2D arrays with high-speed readout have been produced. Monolithic integration with III–V, e.g. GaAs, electronic circuitry is a desired goal. However, integration with silicon circuitry is more common.
- MEMS. Previously introduced in section ‘One-dimensional mechanical modulators and application, large 2D arrays have been developed, e.g. $> 750 \times 500$ pixels for display purposes, but the DMD has generally fewer pixels but higher frame rate. These are strictly E-SLMs and therefore the addressing circuitry limits the frame rate. However, MEMS arrays can be constructed of silicon, so the modulator and addressing circuitry can be monolithically integrated.

SLM functions and goals for optical processing

A large variety of SLM devices were categorized in section ‘Two-dimensional SLM devices categories’. However, to understand the desired functioning of an SLM a general construct can be used, describing an SLM as a ‘sandwich’ of structures (figure 3.5.9): a modulating material between control and conversion structures. The conversion structures handle the input to and output from the SLM, the ‘write’ beam and the ‘read’ beam, respectively, and may consist of photosensitive material for an O-SLM, or an array of electrodes for an E-SLM. The control structure is the third port of this ‘three-port’ device. Figure 3.5.9 can be used to illustrate the variants of SLMs and the basic functions they provide for optical processing:

- Electrical-to-optical transduction. The write image is in electrical form, e.g. the SLM input is via an electron beam, as in a cathode-ray tube display, or via write lines on an electrical circuit. The input information is modulated onto an optical read beam (no read image data).
- Optical-to-electrical transduction. If the write image is in optical form, there is no read image, and the Modulated Output is electrical, then the SLM acts as a classical photodetector array.
- Light conversion. The write image may be on an optical beam of either incoherent light or coherent light of a particular wavelength. The read image is a readout beam of coherent light of another wavelength onto which the write image is modulated to produce an optical modulated output beam. For incoherent optical input, the image data is converted from incoherent to coherent light; in the coherent input case the image data can be converted from one wavelength to another. Additionally, the polarization state of the output can be changed relative to the input polarization.
- Imagery projection. The write image data may be either electrical or optical. If the read image is a readout beam, an optical beam of sufficient intensity then that beam may be used for projection display of the write image data modulated onto it.
- Image processing. An SLM can be used as an input and processing device for image-processing systems that perform operations such as transformations (e.g. Fourier transform), image correlation, nonlinear operations such as thresholding and level slicing, and linear re-mapping of images. The write image data may be either electrical or optical. The read image and the modulated output can then be either optical or electrical, respectively (depending on the write image). Some image processing may be performed first by the SLM, but the modulated output will be into either a further optical or electronic image-processing system.

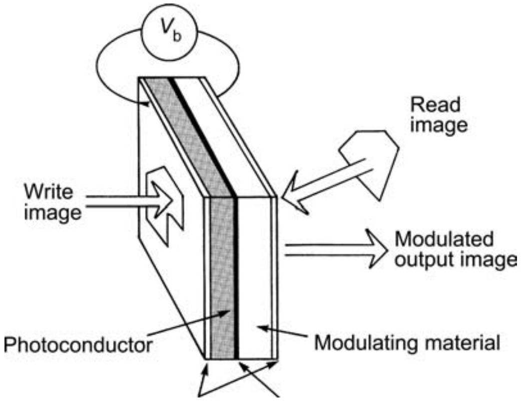


Figure C3.5.9. Generic sandwich construction for a 2D SLM.

- Image amplification. The read image is a readout beam that is modulated to produce an optical modulated output. The optical power in the read beam may be larger than in the write beam, or the SLM can amplify the write beam signal.
- Analogue multiplication, addition and subtraction of data arrays. Data on the write beam can interact with data on the read beam using any of a variety of mechanisms to perform these operations.
- Storage of data arrays. The modulating material may retain the information impressed onto it for some time. During the retention time a modulated output is obtained. If the SLM is optically read out, the modulating material is an optical memory.
- Input-to-output isolation and gain element for 2D data arrays, akin to a transistor for 1D temporal data. This three-port behavior is particularly important for optical interconnection and cascading optical elements in an all-optical architecture, since with a three-port device the modulated output levels are independent of the strength of the write image signal.
- Analogue phase conjugator for adaptive optical systems. Write and read images and modulated output are all optical; the SLM performs the phase conjugation operation.

For optical information processing systems utilizing 2D SLMs to be effective, SLM devices need to exhibit a minimum throughput (frame rate times pixels/frame) on par with that already demonstrated with acousto-optic devices, roughly 10^9 samples/s. However, this level of throughput can be partitioned between frame rate and frame size, as in the following two examples.

- A liquid crystal display of 10^7 pixels and 30 frames/s would approach the 10^9 samples/s goal.
- MQW modulator with 6.5×10^4 pixels (256×256 device) at 10^5 frames/s would exceed the goal. Such a modulator would likely be optically addressed, to avoid use of ultra-high speed electronic addressing circuits.

Depending on application it is likely that 10^9 pixels/frame would be desirable for processing of large images, and $>10^3$ framing rate would be needed for applications such as 3D displays and rapid search of volume holographic databases (see section 'Optical storage'). In addition to these two parameters, additional requirements are:

- Large dynamic range and good image quality. A uniform 1000:1 contrast ratio would satisfy most requirements.
- Low power consumption. A power consumption figure of merit would have units of $\mathrm{mW\,cm^{-2}\,kHz^{-1}}$. This figure of merit would relate also to the size and resolution of the SLM. To keep power consumption low for a given number of pixels, the size of the pixels needs to be minimized. Pixellation corresponding to 50 line pairs/mm, easily within the state-of-the-art of most fabrication technologies, would minimize the focal length of the optics needed for relaying information to/from the SLM, and overall minimize optical system size and power.
- Data storage time. Flexibility to address a variety of applications (section 'Application of 2D SLM devices') is desirable, e.g. a fast frame rate SLM with storage times of seconds to almost indefinite. It may be necessary to keep the same data displayed in an SLM indefinitely as one rapidly processes imagery or searches databases. Another example, would be where the SLM responds to slowly-changing phenomena, such as environmental conditions, in an adaptive optics application.

> At the extreme, the data are time invariant; except for write-time considerations, use of film as an SLM would be acceptable. Time-invariant applications include feature extractors and fixed filters for image processing.

Applications of 2D SLM devices

Image processing

Next to the use of optical processing in synthetic aperture radar image formation, the 2D optical processor with the longest history is the Fourier-transform based image correlator. This processor has arguably seen the most development over a number of years, with interest for high speed automatic target recognition systems. But performance has been determined by available SLM devices and photodetector arrays. Because of a lower requirement on space-bandwidth than alternate architectures, and ability to use direct phase encoding for the input and filter function, the preferred correlator architecture is a modification of the arrangement for temporal correlation, figure C3.5.8. Known as the VanderLugt correlator, 2D SLMs are used at the input image plane $f(x,y)$ and the Fourier-filter plane, and a 2D photodetector is used to output the correlation function (at the location of the second acousto-optic cell in figure C3.5.8, and the last lens and point photodetector are eliminated). Correlation of one image $f(x,y)$ with another $s(x,y)$ is performed as multiplication in Fourier space followed by inverse Fourier transformation. The key to use of this architecture is the scheme for encoding both phase and amplitude of the Fourier-plane filter onto an SLM device or film [63]. Variants of the VanderLugt filter have been explored to match the capabilities of existing SLMs and to improve throughtput. These variants generally take advantage of the well-known fact in signal processing that clipping the amplitude of a signal does not greatly degrade the detection margin in the correlation process, only a 1 dB penalty typically. Hence, binary amplitude Fourier filters have been used. A further extension has been to additionally binarize the phase values in the filter, i.e. a binary phase only filter where the amplitude transmittances of ± 1 correspond to phase shifts of 0 and π. Various high-frame-rate SLMs with limited or no grey scale capability are therefore viable candidates for a high-speed correlation system such as the FLC SLM, the DMD SLM, and the electro-optic PLZT-on-silicon SLM. A compact correlator operating at 500 Hz using FLC technology has been packaged in a volume of $0.6' \times 1.0' \times 0.4'$ [15].

Many issues in producing viable optical processors have been addressed for optical correlators. The performance and size of an optical correlator is found to depend on SLM pixel pitch and size, and device flatness and uniformity. For example, the level of detail of the Fourier-plane filter will increase with number of pixels in the SLM and thus increase the ability to differentiate among various objects. Also, the correlator system size scales linearly with the number of pixels and quadratically with pixel pitch [15]. Thus, smaller pixel pitch in larger more-uniform frames are needed. However, equally important are the performance parameters of the photodetector array. Since it is often only necessary to detect the location of the peak of the correlation function, many proposals have been made for smart photodetector arrays that can report correlation results as peak/no peak at rates commensurate with the SLM rate. If peak detection is insufficient for the application, then faster image processing will be required to extract correlation plane features. Whether such processing can be done in another stage of processing remains a future challenge. Unless many of these issues are addressed, optical systems will remain noncompetitive with all-electronic implementations.

Adaptive techniques are a way to improve the performance of optical processing systems, and SLM technology can been applied to such techniques. One technique is to perform phase conjugation on an optical wavefront to restore the wavefront to its original state by removing aberrations and distortions introduced by the environment or imperfections in an optical system. Phase conjugation requires the storage of detailed interference fringe patterns produced by the aberrated wavefront and an unaberrated reference beam. A phase-conjugate mirror can be implemented with an SLM constructed of a photorefractive material [64]. Progress to date has centred on development of the required material parameters. High speed of response and efficiency are required from the photorefractive material. Photorefractive quantum-well material is capable of sub-microsecond response, and the efficiency has been increased by constructing pixels consisting of etalons [3]. Phase conjugate efficiency of 100% with high net gain has been achieved with polymer photorefractive materials [65].

Optical interconnection

Free-space: crossbar switch. Free-space optical interconnection can be described in matrix formulation that is easily related to capabilities in optical systems. Consider an array of optical input and output data channels, such as fibre-optic. If the array of input channels is represented as a vector $\boldsymbol{u}$, any element or combination of elements can be routed to any element of the output array or vector $\boldsymbol{v}$ using an appropriately constructed matrix $\boldsymbol{M}$ in the equation

$$\boldsymbol{v} = \boldsymbol{M}\boldsymbol{u}. \tag{C3.5.20}$$

SLMs are needed in this concept because of several desirable characteristics they can provide. First, when interconnecting among a number of different communications channels an SLM provides isolation between input and outputs. Second, the routing process is naturally lossy, because each element of $\boldsymbol{v}$ must fan out onto a row of $\boldsymbol{M}$. Optical-to-optical SLMs can then provide the amplification gain to compensate for these optical losses.

Optical storage

Effective processing systems need to have access to memory. It is natural to explore possible use of optical processing hardware with optical storage technology. One function of an SLM is storage of data, as noted in section ‘SLM functions and goals for optical processing’. An SLM can store the results from one stage of processing, and present those results to the next stage. However, the amount of information that can be stored on an SLM is limited to the number of pixels in the device. Optical devices that store large amounts of data are, however, of interest in optical information processing, e.g. large databases may need to be accessed rapidly and repeatedly, such as for cross-correlation. Conventional high-density optical storage technology has been reviewed in chapter C3.4. However, conventional optical disk storage technology has not been found useful for optical processing without significant modification or additional electronics. One nonconventional approach stores a library of analogue holograms on a rotating photopolymer disk for subsequent correlation. Large numbers of holograms can be stored on a single disk, and rapid scanning of the library is possible with standard mechanical rotary drives. Another disk approach is to store conventional digital data in a 2D page format. Such holographic disk systems have been demonstrated that perform at 1 Gbit s^{-1} sustained and 10 Gbits s^{-1} peak transfer rates; along with the holography, is electronic processing that includes data modulation coding, interleaving, and error correction to a bit error rate of 10^{-3} [66].

Volume holographic storage has a more natural optical interface using 2D SLMs to transfer data from previous and to subsequent stages of 2D optical processing. Volume storage also would have a much higher storage capacity than disk, with a theoretical storage limit of ~ 10 Tbits cm^{-3} ($\sim$ volume/λ^3 where λ is approximately the smallest linear dimension for storing one bit). In comparison, present single-layer optical disks have areal storage density of 15 Gbits in^{-2}, which can be extended with multilayer disks, the increase depending on the number of layers and the laser wavelength. Among the issues in cascaded optical SLM and storage elements is rapid addressing of and access to data arrays, and the proper registration of data arrays. The latter issue has been addressed with electronic post-processing techniques [67].

C3.5.3 Summary

The area of optical information processing has explored many avenues of research. Areas of application are still developing, mainly as devices capabilities increase and new capabilities emerge. Many lessons have been learned on the proper application of optical capabilities.

- Using optics for general numeric computations is not presently desirable, because many of the digital operations are intrinsically nonlinear. Necessary nonlinear optical devices are still large compared to electronic integrated circuits, and also inefficient, i.e. power hungry. This does not mean the architectures that have been developed are without value. Such architectures were developed mainly to cleverly exploit the parallel-processing capabilities of optics, and may be applicable to building specialized processors [68].
- The advent of optical communications has led to a near-term processing application for fibre-optic interconnects between high-performance electronic digital processors and sensor arrays, whereby simple operations are performed on the optical data within fibres, such as multiplexing/ de-multiplexing and signal conditioning. Fibre interconnects are a straightforward insertion of optics, mainly to address input–output data-bandwidth bottlenecks that can occur with high-speed digital processing elements. But developments in fibre-optic technology have important implications for optical processing; these include the development of wavelength diversity techniques, and of all-optical techniques, presently used to cascade fibre-optical subsystems in a network to avoid costly conversions between the optical and electrical domains. However, with fibre optics the full free-space 3D interconnection capability of optics is not fully exploited. Full exploitation of free-space 3D interconnection technology will require the use of wavelength diversity and all-optical techniques employing cascaded 2D SLMs. There will also be concurrent development of novel multi-processor computing architectures. Candidate architectures are likely to involve neural network parallel-distributed-processing concepts, where the interconnection scheme can be integral to the processing.
- Analogue optical signal processors, particularly for Fourier-transform-based operations such as spectral analysis and correlation, have been developed because they can perform linear operations, at a given rate, in implementations with smaller size-weight-power product than all-electronic implementations. Avoidance of power-hungry analogue-to-digital conversion by processing directly in the analogue domain and concentrating on applications in the microwave-frequency region make such processors a true complement to digital processors. Most development of compact systems has been done with 1D acousto-optic modulators, due to the maturity of acousto-optic technology. However, the highest performance should be obtained using 2D SLMs addressing intrinsically 2D problems such as pattern recognition and 3D optical interconnection. Use of wavelength-diversity techniques should further increase performance. Various 2D SLMs are capable of implementing functions needed for all-optical architectures, especially optical amplification, data storage, and data thresholding. All-optical systems are needed to address the fundamental problem with present analogue processors, by optically performing the present electronic pre- and post-processing of data into/out of optical processors. Such processing includes dynamic range re-scaling and data formatting, various nonlinear operations, and signal amplification and data encoding schemes. These needs can be met potentially with developments in optical-to-optical SLMs, adaptive optics, and volume holography. Throughput improvements of 2D devices are still required, since present

throughput rates are lower than for the acousto-optic devices. Hence, initial implementations may still be restricted to 1D and quasi-multi-dimensional architectures.

A feasible vision for the future for all application areas therefore involves all-optical architectures. Optical device technology exists for implementing subsystems that perform optical input, processing, and storage functions. The challenge is to develop efficient information and data flow between such subsystems in the optical implementations of pre- and post-processing functions that are done electronically for present optical systems.

References

[1] Gopalakrishnan G K, Burns W K, McElhanon R W, Bulmer C H and Greenblatt A S 1994 Performance and modeling of broadband $LiNbO_3$ traveling wave optical intensity modulators *J. Lightwave Technol.* **12** 1807

[2] Noguchi K, Mitomi O and Miyazawa H 1998 Millimeter-wave Ti: $LiNbO_3$ optical modulators *J. Lightwave Technol.* **16** 615

[3] Bowman S R, Rabinovich W S, Beadie G, Kirkpatrick S M, Katzer D S, Ikossi-Anastasiou K and Adler C L 1998 Characterization of high performance integrated optically addressed spatial light modulators *J. Opt. Soc. Am. B* **15** 640

[4] Goodman J W 1968 *Introduction to Fourier Optics* (New York: McGraw-Hill)

[5] Athale R A, Szu H H and Lee J N 1981 Three Methods for Performing Hankel Transforms *Optical Information Processing for Aerospace Applications* NASA Conference Publ. 2207, NASA CP-2207 133

[6] Anderson G W, Webb D, Spezio A E and Lee J N 1991 Advanced Channelization Devices for RF, Microwave, and Millimeterwave Applications (Invited Paper) *Proc. IEEE 79*, pp 355–388, 1991

[7] Griffin R D and Lee J N 1996 Acousto-optic wide band correlator system *Acousto-Optic Signal Processing* 2nd edn chapter 11, ed N Berg and J Pelligrino (New York: Dekker) pp 367–400

[8] Berg N J and Lee J N (eds) 1983 *Acousto-optic Signal Processing: Theory and Implementation* (New York: Dekker)

[9] Huang A 1978 An optical residue arithmetic unit *Proc. Fifth Annual Symposium on Computer Architecture* (New York: Association for Computing Machinery and IEEE Computer Society) pp 17–23

[10] Whitehouse H J and Speiser J M 1977 *Aspects of Signal Processing with Emphasis on Underwater Acoustics*, ed G Tacconi (Hingham, MA: Reidel)

[11] Athale R A, Hoang H Q and Lee J N 1983 High accuracy matrix multiplication with magnetooptic spatial light modulator *Real-Time Signal Processing VI: Proc. SPIE* vol 431, ed K Bromley p 187

[12] Miller D A B, Chemla D S, Damen T C, Gossard A C, Wiegmann W, Wood T H and Burrus C A 1984 Band-edge electroabsorption in quantum well structures: the quantum confined Stark effect *Phys. Rev. Lett.* **53** 2173

[13] Lentine A L, Hinton H S, Miller D A B, Henry J E, Cunningham J E and Chirovsky L M F 1989 Symmetric self-electro-optic effect device: Optical set-reset latch, differential logic gate, and differential modulator/detector *IEEE J. Quantum Electron.* **25** 1928

[14] Yu S and Forrest S 1993 Implementations of smart pixels for optoelectronic processors and interconnection systems: I. Optoelectronic gate technology, and II. SEED-based technology and comparison with optoelectronic gates *J. Lightwave Technol.* **11** 1659

[15] Turner R M, Johnson K M and Serati S 1995 High speed compact optical correlator design and implementation *Design Issues in Optical Processing*, ed J N Lee (Cambridge: Cambridge University Press) p 169

[16] Denker J S ed 1986 *Neural Nets for Computing* (New York: American Institute of Physics)

[17] Psaltis D and Farhat N 1985 Optical information processing models based on an associative-memory model of neural nets with thresholding and feedback *Opt. Lett* **10** 98

[18] Fisher A D, Lippincott W L and Lee J N 1987 Optical implementations of associative networks with versatile adaptive learning capabilities *Appl. Opt.* **26** 5039

[19] McCormick F B and Tooley F A P Optical and mechanical issues in free-space digital optical logic systems *Design Issues in Optical Processing*, ed J N Lee (Cambridge: Cambridge University Press) p 220

[20] Lentine A L *et al* 1997 ATM distribution network using an optoelectronic VLSI switching chip *Optics in Computing, OSA Technical Digest* **8** 2

[21] Husain A, Crow J and Lee J N (eds) 1991 *J. Lightwave Technol: Special Issue on Optical Interconnects*;
Husain A and Lee J N, (eds) 1995 *Special Issue on Optical Interconnects*.

[22] Feldman M R, Camp J L, Sharma R and Morris J E 1995 Comparison between holographic and guided wave interconnects for VLSI multiprocessor systems *Design Issues in Optical Processing*, ed J N Lee (Cambridge: Cambridge University Press)

[23] Uomi K, Mishima T and Chinone N 1985 Ultra-high relaxation oscillation frequency (up to 30 GHz) of highly p-doped GaAlAs multiquantum well lasers *Appl. Phys. Lett.* **51** 78

[24] Yariv A 1989 *Quantum Electronics* 3rd edn. (New York: Wiley) chapters 14 and 16

[25] Hecht D L 1977 Spectrum analysis using acousto-optic filters *Opt. Eng.* **16** 461

[26] Korotsky S K and DeRidder R M 1990 Dual parallel modulation schemes for low-distortion analogue optical transmission *IEEE J. Sel. Areas Commun.* **8** 1377

[27] Johnson L M and Rousell H V 1988 Reduction of intermodulation distortion in interferometric optical modulators *Opt. Lett.* **13** 928

[28] Walker R G 1991 High-speed III–V semiconductor intensity modulators *IEEE J. Quantum Electron*. **27** 654

[29] Trezza J A, Pezeshki B, Larson M C, Lord S M and Harris J S 1993 High contrast asymmetric Fabry–Perot electroabsorption modulator with zero phase change *Appl. Phys. Lett.* **63** 452

[30] Farnett E C, Howard T B and Stevens G H 1970 Pulse compression radar *Radar Handbook* Chapter 20, ed M I Skolnik (New York: McGraw-Hill)

[31] Frankel M Y, Matthews P J and Esman R D 1996 Two-dimensional fibre-optic control of a true time-steered array transmitter *IEEE Trans. Microwave Theory and Techniques* **44** 2696

[32] Wong Y M *et al* 1995 Technology development of a high-density 32-channel 16 Gb/s optical data link for optical interconnection applications for the Optoelectronic Technology consortium (OETC) *J. Lightwave Technol.* **13** 995

[33] Chang-Hasnain C J, Maeda M W, Harbison J P, Florez L T and Lin C 1991 Monolithic multiple wavelength surface emitting laser arrays *J. Lightwave Technol.* **9** 1665

[34] Lau K Y, Derry P L and Yariv A 1988 Ultimate limit in low threshold quantum well GaAlAs semiconductor lasers *Appl. Phys. Lett.* **52** 88

[35] Rabinovich W S *et al* 2001 InGasAs multiple quantum well modulating retro-reflector for free space optical communications *Patent pending*, U.S. Naval Research Laboratory 2000

[36] Klein W R and Cook B D 1967 Unified Approach to Ultrasonic Light Diffraction *IEEE Trans. Sonics Ultrasonics*

[37] Phariseau P 1956 *Proc. Indian Acad. Sci.* **44A** 165–170

[38] Uchida N and Niizeki 1973 Acousto-optic deflection materials and techniques *Proc. IEEE* **61** 1073

[39] Lee J N (ed)1995 *Design Issues in Optical Processing* (Cambridge: Cambridge University Press)

[40] Young E H and Yao S K 1981 Design considerations for acousto-optic devices *Proc. IEEE* **69** 54

[41] Cohen M G 1967 Optical study of ultrasonic diffraction and focusing in anisotropic media *J. Appl. Phys.* **38** 3821

[42] VanderLugt A, Moore G S and Mathe S S 1983 Multichannel Bragg cell compensation for acoustic spreading *Appl. Opt.* **22** 3906

[43] Amano M and Roos E 1987 32-channel acousto-optic Bragg cell for optical computing *Acoustooptic, Electrooptic, and Magnetooptic Devices and Applications* **753** 37

[44] Anderson G W, Guenther B D, Hynecek J A, Keyes R J and VanderLugt A 1988 Role of photodetectors in optical signal processing *Appl. Opt.* **27** 2871

[45] Turpin T M 1978 Time integrating optical processing *Proc. SPIE: Real-Time Signal Processing I* **154** p 196

[46] Lee J N 1987 Architectures for temporal signal processing *Optical Signal Processing*, ed J L Horner (New York: Academic) p 165

[47] Turpin T M 1981 Spectrum analysis using optical processing *Proc. IEEE* **69** 79

[48] Cohen J D 1979 Ambiguity processor architectures using one-dimensional acoustooptic transducers *Real-Time Signal Processing II: Proc. SPIE* vol 180 p 134

[49] Athale R A and Lee J N 1983 Optical systems for efficient triple-matrix-product processing *Opt. Lett.* **8** 590

[50] Haney M and Psaltis D 1985 Acousto-optic techniques for real-time SAR imaging *Proc. SPIE: Optical Technology for Microwave Applications II* **545** 108

[51] Lee J N and VanderLugt A 1989 Acousto-optic signal processing and computing *Proc. IEEE* **77** 1528

[52] Essex Corp 2000 *ImSym Product Specification*.

[53] Johnson R V, Hecht D L, Sprague R A, Flores L N, Steinmetz D L and Turner W D 1983 Characteristics of the linear array total internal reflection electro-optic spatial light modulator for optical information processing *Opt. Eng.* **22** 665

[54] Hecht D L 1993 Advanced optical information processing with total internal reflection electro-optic spatial light modulators *Proc. SPIE: Optical Information Processing* **2051** 306

[55] Dixon R W 1967 Acoustic diffraction of light in anisotropic media *IEEE J. Quantum Electron.*

[56] Smith D A, Baran J E, Cheung K W and Johnson J J 1990 Polarization-independent acoustically tunable filter *Appl. Phys. Lett.* **60** 1538

[57] Willner A E and Smith D A 1996 Acousto-optic modulators flatten amplifier gain *Laser Focus World* 177 June

[58] Young E H and Belfatto R V 1993 Polychromatic acousto-optic modulators let users tailor output wavelengths *Laser Focus World* p 179 November

[59] Spring K R 2002 Acousto-optic tunable filters improve optical microscopy *Laser Focus World* January p 123

[60] Pepper D M, Feinberg J and Kukhtarev N V 1990 The photorefractive effect *Sci. Am.* 62

[61] White, J O, and Yariv, A 1984 Photorefractive crystals as optical devices, elements, and processors *Solid State Optical Control Devices: Proc SPIE*, vol 464.

[62] Agranat, A J 2002 Electroholographic switching devices and applications *2002 CLEO Conference Technical Digest* (Washington: Optical Society of America) p 37 Paper CMH1

[63] VanderLugt A B 1964 Signal detection by complex spatial filtering *IEEE Trans. Inform. Theory* **10** 139

[64] Grunnet-Jepsen A, Thompson C L and Moerner W E 1997 Spontaneous oscillation and self-pumped phase conjugation in a photorefractive polymer optical amplifier *Science* **277** 549

[65] Peyghambarian N, Meerholz K, Volodin B, Sandolphon and Kippelen B 1994 Near 100% diffraction efficiency and high net-gain in a photorefractive polymer *Optics and Photonics News* December p 13

[66] Orlov S S, Phillips W, Bjornson E, Hesselink L and Okas R 2002 Ultra-high transfer rate high capacity holographic disc digital data storage system *Proc 29th Applied Imagery Pattern Recognition Workshop* (New York: IEEE) p 71

[67] Burr G W and Weiss T 2001 Compensation of pixel misregistration in volume holographic data storage *Opt. Lett.* **26** 542

[68] Tippett J T, Berkowitz D A, Clapp L C, Koester C J and Vanderburgh A Jr. 1965 *Optical and Electro-optical Information Processing* (Cambridge, MA: MIT)

[69] Brooks R E 1985 Micromechanical light modulator on silicon *Opt. Eng.* **24** 101

[70] Hornbeck L J 1989 Deformable-mirror spatial light modulators *Proc. SPIE Spatial Light Modulators and Applications III* **1150** 86

[71] Chang I C 1974 Noncollinear acousto-optic filter with large angular aperture *Appl. Phys. Lett.* **25** 370

C4.1
Spectroscopic analysis

Günter Gauglitz and John Dakin

C4.1.1 Introduction

Optical spectroscopy is an invaluable tool for the characterization of many physical and chemical components and processes. As in so many other areas, where science or engineering can offer benefits, nature has already evolved optical spectroscopy as a very useful tool. Our eyes are our primary sensing means and cannot only determine the range of remote objects and observe their fine spatial details, but they can also use spectroscopy to detect their colour, albeit with rather poor spectral resolution. However, modern science can, using sophisticated hardware, determine far more than is possible with simple visual inspection. It can provide much more complex and quantitative data and use it to characterize, precisely, a wide variety of physical objects or chemical analytes. Instruments cannot only achieve much higher spectral resolution, but can also cover a much wider wavelength range and, if desired, even provide time- or distance-resolved results. A major advantage of optical methods of analysis, is that they are usually non destructive and non invasive, and usually quick in operation.

In optical spectroscopy, electromagnetic radiation in the range of $10^{12}-10^{15}$ Hz is most commonly used, covering infrared, visible and ultraviolet radiation. Optoelectronic devices detect this electromagnetic radiation by means of its interaction with matter.

This section will review the main methods of spectroscopy, concentrating mainly on optical aspects, adding further detail to earlier sections, where needed. It will also briefly review, and where appropriate extend, the earlier treatment of interactions of radiation with matter and the transport of radiation via fibre optics or waveguides. The section is effectively structured in two halves, the first covering general concepts of spectroscopy and the second describing a number of illustrative case studies involving spectroscopy, sometimes with fibre optics, mostly from the research team of Prof. Gauglitz, in Tuebingen, Germany.

The passage of electromagnetic radiation can be influenced by matter in many different ways, with processes such as: refraction, polarization change, specular reflection, scattering and absorbance [1] being just a few examples. It should be noted that the magnitude of many of these effects depends significantly on optical wavelength, or photon energy.

The technique of absorbance spectroscopy, or spectrophotometry, monitors attenuation of a beam of light by matter and is commonly applied in spectrometry, particularly for the infrared, visible and ultraviolet regions.

Measurement of light lost from a beam due to scattering (*nephelometry* or *turbity* measurement) is also often used. Apart from the more common light-scattering mechanisms, vis. Rayleigh, Rayleigh–Gans, Mie and geometric scattering, other elastic scattering (i.e. scattering with no change of wavelength) can arise due to the effects of diffraction from periodic structures (i.e. like x-ray scattering in crystals).

Inelastic scattering, which involves a change in optical wavelength (known as inelastic, because changes in photon-energy occur) can also arise, for example in Raman or Brillouin scattering, where a photon-phonon interaction process occurs.

After absorption of light, matter can dissipate the absorbed energy by several means, for example from an excited electronic state to lower vibrational or rotational states, i.e. to produce phonons or heat. The absorbed radiation can also be re-radiated, resulting in luminescence behaviour (this behaviour can be sub-divided into fluorescence and phosphorescence). In addition, after excitation with particularly energetic particles, matter can re-emit strong radiation, a behaviour made use of in x-ray emission spectroscopy and atom emission spectroscopy.

In some cases, for example in flames, thermally-excited electrons, in the hot gaseous materials, can exhibit direct radiation of light by energy transitions from the excited states.

Although an introduction to electromagnetic radiation and its properties was given earlier in Chapter A1.1, most of these processes we mentioned above will be briefly discussed again below [2–5], with a few additions relevant to spectroscopy. We shall add a few relevant points and concentrate particularly on the UV/VIS/NIR spectral region, especially when discussing quantitative aspects. Here, the theoretical principles of atomic states and dispersion will be discussed, ending with some brief photophysics and a description of certain aspects of polarization states in waveguides. A description of several important spectroscopic methods and types of spectrometers will be given, concentrating on practical aspects, and the section will end with a brief description of some interesting case studies of optical sensor systems, with special reference to evanescent field techniques.

C4.1.2 Theoretical principles

C4.1.2.1 Properties of electromagnetic radiation

Electromagnetic radiation and its interaction with matter were discussed in detail in Chapters A1.1 and A1.2, so the properties of radiation will only be summarized briefly here. Electromagnetic radiation is characterized by its amplitude, frequency, state-of-polarization, phase and the time dependence of its amplitude. Depending on the physical effects to be described, electromagnetic radiation can best be considered as either a particle or a wave, a behaviour called *de Broglie dualism*. Choosing to consider it as a particle, or *photon*, gives a good physical explanation of phenomena involving effects of impact or momentum, i.e. photo-emission (or Compton effect). The choice of the alternative wave description, however, is best used to understand interference and refraction effects [6, 7]. The impact momentum, p, is proportional to the inverse of the wavelength, λ, of the radiation according to equation C4.1.1:

$$p = \frac{h}{\lambda} \tag{C4.1.1}$$

This can be given, in terms of mass, m, and velocity, v, by:

$$p = mv \tag{C4.1.2}$$

p is the momentum, given in $\mathrm{kg\,m^{-1}s^{-1}}$ and h is Planck's constant $= 6.63 \times 10^{-34}\,\mathrm{J}$. In the second equation, m is the effective mass in kg, and v is the velocity in $\mathrm{ms^{-1}}$. Propagation of electromagnetic radiation can be explained by the well-known Maxwell's equations, which predict that, in a propagating E–M wave, the electric and the magnetic field vectors lie perpendicular to each other and are perpendicular to the direction of propagation (see figure C4.1.1). Considering the particulate (or quantized) nature of the light, the radiation is composed of a stream of photons and the energy of each is

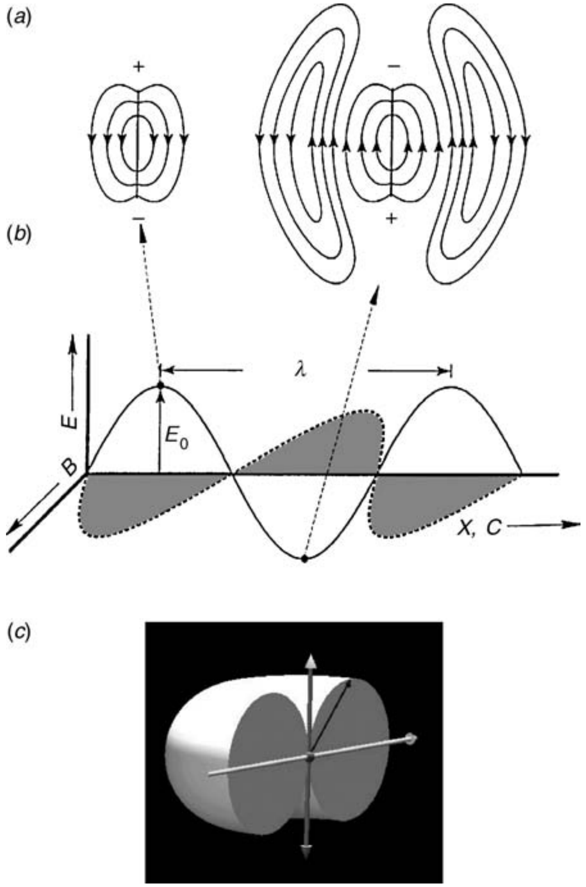


Figure C4.1.1. Electrical fields of a Hertz dipole. These are perpendicular to both the magnetic field and to the direction of propagation of the electromagnetic radiation (*a*). The time dependence of the electric field distribution is given in (*b*). The polar diagram of radiated intensity (power) is shown in (*c*).

proportional to the frequency in the wave model, so is therefore reciprocally dependent on the optical wavelength (see equation C4.1.3).

$$E = h\nu = \frac{hc}{\lambda} = hc\tilde{\nu} \tag{C4.1.3}$$

Here $c = 3 \times 10^8\,\mathrm{ms}^{-1}$, is the velocity of light in vacuum, and $\tilde{\nu}$ is the wavenumber in cm^{-1}.

The unit of optical *frequency* that is most commonly used by scientists working in the IR region, or Raman spectroscopists, is the *wavenumber* and in these research fields it is in very general use, despite clearly not being an S.I. unit. In the field of highly energetic radiation, eV (electron volts) is usually used. In the UV (ultraviolet), VIS (visible) and NIR (near infrared) wavelength range, most scientists use units of wavelength, λ, in nm, or, in the infrared (IR), units in μm. Yet again, many of these are not S.I. units, reflecting the fact that the reasons for the choice of the different units were historical in nature. At least the choice usually has the advantage of avoiding using numbers or fractions that have an inconvenient (too large or too small) size.

The approximate order-of-magnitude ranges of wavelength, frequency, wavenumber and photon energy that correspond to a number of different types of electromagnetic radiation are shown in table C4.1.1 and figure C4.1.2. The typical type of interaction that each type of electromagnetic radiation might have with matter is also shown, but this will be discussed later. As can be seen, the frequency and wavelength scales are logarithmic in the table, in order to cover a very wide range.

Electromagnetic radiation in the ultraviolet and visible and infrared wavelength range is especially interesting for spectroscopy. For this reason, this particular spectral range has been 'zoomed out' in figure C4.1.2 and we have indicated the colours associated with the different parts of the visible spectrum. In the case of simple absorption photometry, a transmission measurement is taken at only one wavelength (i.e. monochromatic light is used), but in more general spectrophotometry, a transmission spectrum is usually measured over a wider wavelength range. Such a spectrum will show transmission changes arising due to not only one type of absorption, but also a more complex absorption behaviour involving perhaps a superposition of many rotational, vibrational and electronic interactions.

C4.1.2.2 Interaction between radiation and matter

Although we run the slight risk of duplicating descriptions in other sections, we shall now, for the convenience of the reader, briefly review how electromagnetic waves react with matter. We shall start with elastic processes; vis. ones that involve no loss (or gain) of energy by the light, and then extend the discussion to cover inelastic processes, where photon energy changes occur.

Dispersion

The most usual and important interaction with matter for use in spectroscopy involves the electric field vector. The periodic change in the electric field vector of an incident E–M wave can cause either a new orientation of a molecule or a displacement of electrical charges in a dielectric medium (matter). Both of these result in electrical *polarization* of the matter, a so-called *dielectric effect* process. If there is any variation of this effect with optical wavelength or frequency, the effect is called *dispersion* [6], as it changes the velocity of the wave.

Radiation induces a type of Hertz dipole in a dielectric material and if this material is not optically homogenous with its surroundings, it can result in radiation, or *scattering*, into these surroundings. The charge displacement in a dielectric in an electromagnetic wave can be explained in terms of a simple harmonic-oscillator model, where an alternating dipole is induced in each molecule or micro-region of the material. The total effect is proportional to the number of molecules or macro-molecular particles affected, i.e. polarization is a volume-based quantity.

This Hertz dipole oscillates at the excitation frequency of the radiation and behaves like a radiation source, emitting radiation most strongly in a direction perpendicular to the direction of polarization, as was shown earlier in figure C4.1.1(*c*). When propagating through matter, the radiation maintains its frequency, because of the conservation of photon energy $h\nu$, but decreases in velocity. The velocity of propagation within matter is inversely proportional to the refractive index, n, given by the ratio between the electric field vectors of the radiation propagating in vacuum and matter:

$$n = \frac{\vec{E}(\mathrm{vac})}{\vec{E}(\mathrm{mat})} \tag{C4.1.4}$$

According to Maxwell's equation, the square of the optical refractive index (i.e. the value of refractive index in the optical range of the dispersion curve) is equivalent to the dielectric constant, which, in turn, can conveniently be described mathematically in terms of an equation with real and imaginary parts. This latter parameter can be graphically depicted, as shown in figure C4.1.3.

Table C4.1.1. Spectral regions of electromagnetic radiation, giving the approximate order of magnitude of relevant wavelengths, frequencies, wavenumbers and energies, plus indications of typical interactions that particular type of radiation might have with matter.

Spectroscopic technique	Approximate order of the wavelength [nm]	Approximate order of the frequency [s^{-1}]	Approximate order of the wavenumber [cm^{-1}]	Approximate magnitude of energy (in various units)			Type of interaction
				[kcal mol^{-1}]	[kJ mol^{-1}]	[eV]	
Nuclear magnetic resonance (NMR) 0.1–10 m	10^{10}	3×10^{7}	10^{3}	3×10^{-6}	1.2×10^{-5}	1.2×10^{-7}	Excitation of magnetic transitions of nuclei (spin > O)
Electron spin resonance (ESR) 0.1–10 cm	10^{8}	3×10^{9}	10^{-1}	3×10^{-4}	1.2×10^{-3}	1.2×10^{-5}	Excitation of un-paired electrons
Microwaves 0.1–10 cm	10^{6}	3×10^{11}	10	3×10^{-2}		1.2×10^{-3}	Excitation of rotation of molecules
IR 0.78–10^{3} μm, 3000–10^{6} nm	10^{4}	3×10^{13}	10^{3}	3	12	1.2×10^{-1}	Absorption by excitation of molecular vibrations. Raman scattering
VIS 380–780 nm; UV 200–400 nm; UVU 100–200 nm	10^{2}	3×10^{15}	10^{5}	3×10^{2}	1.2×10^{3}	12×10^{4}	Excitation of electronic transitions, emission and atomic absorption processes. Raman scattering
x-rays 0.01–10 nm	1	3×10^{17}	10^{7}	3×10^{4}	1.2×10^{5}	1.2×10^{3}	Excitation and transitions of electrons involving inner-orbit energy levels
Moessbauer 100 keV; γ absorption	10^{-2}	3×10^{19}	10^{9}	3×10^{6}	1.2×10^{7}	1.2×10^{5}	Resonant absorption by nuclei
γ radiation; >1 MeV	10^{-4}	3×10^{21}	10^{11}	3×10^{8}	1.2×10^{9}	1.2×10^{7}	Nuclei transformation

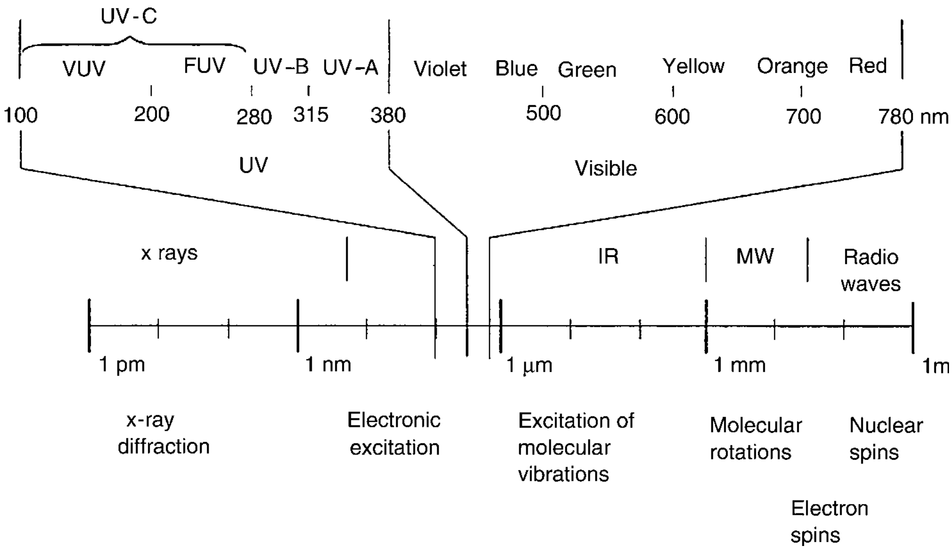

UV-C
VUV
FUV
UV-B
UV-A
Violet
Blue
Green
Yellow
Orange
Red
100
200
280
315
380
500
600
700
780 nm
UV
Visible
x rays
IR
MW
Radio waves
1 pm
1 nm
1 μm
1 mm
1m
x-ray diffraction
Electronic excitation
Excitation of molecular vibrations
Molecular rotations
Nuclear spins
Electron spins

Figure C4.1.2. Depiction of the full electromagnetic spectrum of radiation.

In addition to polarization of matter, as occurs with scattering, absorption of light can also occur. Except in condensed matter (i.e. solids or liquids where spectral broadening occurs due to degeneracy of energy levels) absorption is usually not of a broadband nature, i.e. it often occurs at specific wavelengths. For many materials, e.g. liquids in the microwave region, a reorientation and a relaxation effect can be observed, which changes the effective dielectric constants and causes absorption.

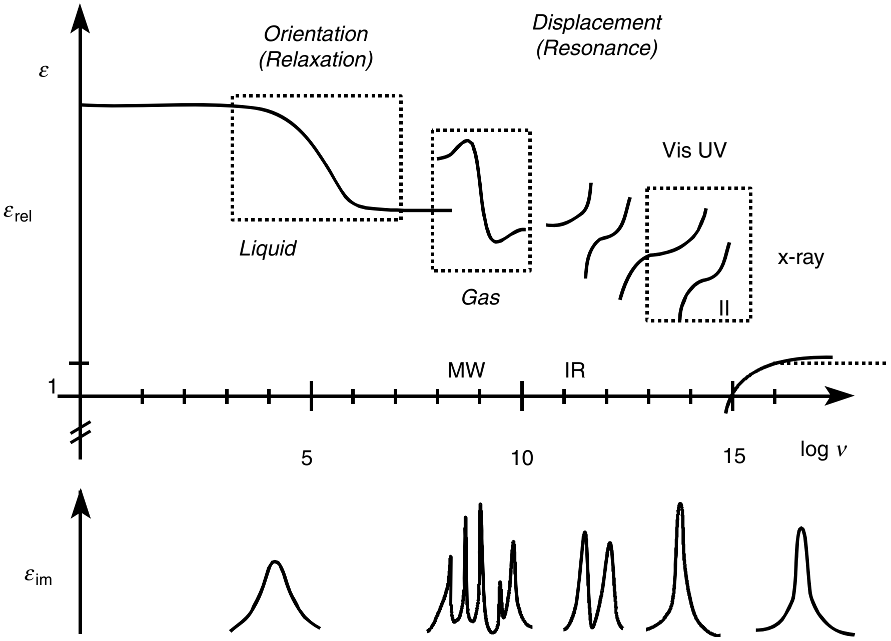

ε
ε_{rel}
Orientation (Relaxation)
Displacement (Resonance)
Liquid
Gas
Vis UV
x-ray
II
1
MW
IR
5
10
15
log ν
ε_{im}

Figure C4.1.3. Depiction of optical dispersion, showing the real (refraction) and imaginary part (absorption) of dielectric constant versus frequency on a logarithmic scale. The lower curve shows various absorption bands.

A resonance (caused by displacement of charges) can typically be observed in the infrared (vibration) and UV/VIS (electronic transition) spectral regions. This causes strong absorption bands (i.e. a larger imaginary part of the dielectric constant). In the case of gases, rotational resonance is also observable at microwave frequencies.

To summarize the above, two fundamental types of interaction between electric fields and dielectric matter can therefore occur: the first is scattering and the second is absorption. These will both be discussed in more detail in the following sections and we shall start with the case of scattering, where no absorption occurs and light energy is conserved.

Scattering

Scattering from single particles

Within molecules, charges may either be symmetrically distributed or they may have an asymmetric distribution, making the molecule *polar* in nature. In non-polar molecules, the nuclei and the electron density distribution have the same centre of charge distribution. However, in an electrical field, these can change and the positive (e.g. atomic nuclei, or positive ions) and negative particles (e.g. electrons, or negative ions) can be displaced relative to each other.

The periodic change in the electric field vector causes a periodic polarization, in line with the polarization of incident light, and the resulting Hertz dipole can then act as an electromagnetic transmitter. The excited system behaves elastically, emitting energy directly from a non-resonant state, with no energy absorbed. Even the smallest of molecules show such scattering effects, but, for such *Rayleigh* scattering (scattering by molecules or particles having a diameter of order of $\lambda/10$ or less), the intensity of scattered radiation increases in proportion to the 6th power of the length of the induced Hertz dipole [8, 9]. For these small particles, the intensity of the radiation varies as the 4th power of the incident optical frequency (if the particle has low optical dispersion) and the light is scattered at essentially the same frequency as the incident beam. It is this strong dependence on wavelength that causes white light from the sun to be scattered mainly at short wavelengths, giving the appearance of the blue sky on a clear day.

The radial distribution of scattered radiation depends, in general, on the shape of the particle or molecule, its dimensions and on the wavelength of scattered radiation. Taking the simplest example, of an isotropic particle, with a diameter smaller than $\lambda/10$, then it can be shown that polarized incident radiation will result in a scattered intensity distribution (polar diagram) as shown in figure C4.1.4*a*.

The reason for this type of 'doughnut-without-a-hole' polar scattering diagram is because a simple Hertz dipole is induced, as discussed earlier. This emits most energy in a direction perpendicular to the axis of the dipole, but emits zero energy in a direction along the axis of the dipole.

If we now consider the case of many very small particles ($<\lambda/10$) having some anisotropy, for example, many randomly oriented, egg-shaped particles, then many induced Hertz dipoles will occur, each at a different angle depending on the orientation of each particle. With incoherent illumination, the scattered intensities from each will add to give a total scattered intensity. Even if the incident light is polarized the polar diagram of the scattering is no longer as in figure C4.1.4(*a*), but has a new shape as in figure C4.1.4(*b*), where there is now a non-zero scattering intensity along the direction of polarization of the incident light. This phenomenon is known as *depolarization*.

When the incident electromagnetic radiation is randomly polarized (or unpolarized), then a set of Hertz dipoles is generated, having axes in all directions normal to that of the incident field, each depending on the instantaneous polarization direction of the incident light. The intensity of radiation in the forward or backward direction is unchanged, but the radiation field at other angles (e.g. at 90°) is a superposition of the wavelets from all these different polarizations, with of course the contribution

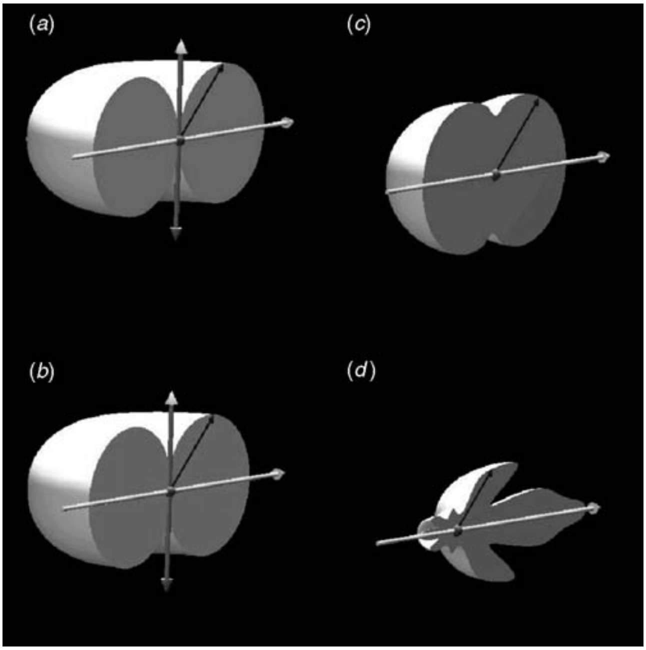


Figure C4.1.4. Polar diagrams for various cases of Rayleigh and Mie scattering. The horizontal arrow shows the direction of incident light. Where the incident light is plane polarized, the vertical arrow indicates the polarization direction. The inclined arrow shows a scattered light direction and the intersection with the outer surface indicates the scattered intensity at this angle. The top left figure (*a*) shows classical Rayleigh scattering, with small particles, where a Hertz dipole is induced along the polarization direction. The bottom left figure (*b*) shows the same when the particles are small, but have a small degree of anisotropy. The top right figure (*c*) shows the case for small particles with randomly-polarized (or unpolarized) incident light. The bottom right figure (*d*) depicts a typical Mie scattering polar diagram for larger particles.

falling to zero whenever the incident polarization (or the induced Hertz dipole axis) is in the direction of observation. This gives a polar diagram with the greatest scattered intensity in the forward and backward directions, but with half that intensity in a direction at right angles to the direction of incident radiation (see figure C4.1.4(*c*)).

If we now take the case where the particle(s) are a little larger (e.g. being of the same order as, or even larger than, the optical wavelength), but where the particle refractive index is very close to that of the surrounding medium, then scattering occurs from many small points within each particle, yet the shape of the incident wavefront remains essentially unchanged during passage through the particle. Then the excitation of each point within the particle is of similar type, but the excitation light has a phase that is dependent on the distance through the particle. This scattering regime is known as *Rayleigh Gans* behaviour. Interference between all the scattered wavelets from different positions will clearly influence the radiation pattern, so, to obtain the far-field polar diagram, the *Rayleigh* scattering intensity from each has to be multiplied by a weighting function, which depends on the geometrical form of the particle [8]. This more complex radiation pattern behaviour is not shown in the diagram, but it also occurs with *Mie scattering*, which is discussed next.

We shall now consider when the scattering centre or molecule is much larger and of significantly higher refractive index compared with its surroundings, Again, more than one centre of scattering within the molecules will arise, but the greater refractive index difference, compared with the surrounding medium, may now be enough to cause wavefront distortion. Thus, there is a phase shift in the transmitted light, relative to light outside the particle. Under these conditions, so-called *Mie scattering* [8] occurs. As with *Rayleigh–Gans* scattering, the scattered intensity distribution is determined by the interference of many scattered wavelets from different sites in the particle, but the situation is now far more complex as the phase of excitation of each point can no longer be so easily determined. The polar diagram is therefore generally complex and usually has many scattering lobes (see figure C4.1.4(*d*)).

The shape of the intensity distribution in the case of all these three scattering types depends on the polarization state of the incident radiation and on the orientation of the scattering molecules. Thus, the intensity distribution and interference pattern contain information regarding the size and the shape of the molecules [8].

It is worth mentioning briefly that there is a further type of scattering, from very large, but optically homogeneous particles (raindrops are a typical example), several orders of magnitude greater than the wavelength of the light, where the scattering can be considered in terms of simple ray tracing, assuming Fresnel refraction and reflection at the optical interfaces. This form of scattering, which is observed in rainbows, is known as *Fresnel* or *geometric scattering*.

Scattering from multiple particles

Scattering from multiple particles depends strongly on whether the light is incoherent (broadband) or coherent (narrow-band). If the light is incoherent, scattered power from each of the particles can simply be summed to give an overall polar far-field scattered intensity that is similar to that from one of the particles, but this is only if not too many strongly scattering particles are present. If the population density of scattering particles becomes very high (as can occur, for example, in milk or in white paints) then the light may suffer strong multi-scattering paths, i.e. substantial secondary, tertiary or even higher order scattering occurs before light from the first excitation (illumination) point can reach the observation point. A medium that causes significant light to be lost by scattering from a direct optical beam passing through it is called a turbid medium. Scattering in such a medium has to allow for this behaviour by taking into account these more complex multiple scattering paths.

If an ensemble of particles is excited with highly coherent light (e.g. monochromatic laser light), the behaviour is very different. Now the scattering from every small particle, or from every part of a larger particle, results in a monochromatic scattered wavelet being re-emitted from each and, at the point of remote observation, the electric fields of all these coherent wavelets will add coherently. This results in constructive, destructive or intermediate levels of interference, depending on the relative phase of the light from the particles, which, in turn, depends on their relative positions relative to both the incident beam and to the direction of, or point of, observation. These effects cause the familiar ‘speckle’ effect when light from a visible laser is first scattered from an object and then the scattered light impinges on a white screen.

In a solid material or particle in solid suspension, the positions of the particles may be reasonably stable, so then true elastic scattering occurs and a relatively-fixed speckle pattern can be observed in the far field scattering diagram, provided, of course, that no significant mechanical movement, vibrations or thermal expansions occur. In fluid (gas or liquid) suspensions, however, the particle will generally undergo more significant and rapid movements due to Brownian motion, convection or turbulence, so the speckle pattern will move continuously with time and because of the small Doppler shifts the light scattering can be described as quasi-elastic. This behaviour, which has come to prominence since the 1970s, can be used to deduce certain useful parameters, such as particle size information, using a method called photon correlation spectroscopy. The light scattered from a monochromatic laser beam interferes at the detector and gives rise to intensity changes (or to changes in photon-arrival-time statistics) which can, with knowledge of the temperature and viscosity of the fluid in which the particles are suspended, be related to the *hydrodynamic radius* of the particles using the Stokes–Einstein equation. The method works best with single-size (monodisperse) particles, but more complex correlation functions from suspensions of two or even three particle types/sizes can be inverted using Laplace transformation.

Reflection

Reflection usually describes the deviation of a light beam such that it doesn't significantly pass through a boundary between two materials, but is re-radiated back into the incident medium. If only single particles (or random ensembles of such particles) are present in gases or liquids, they can be considered to be discontinuities that will cause diffuse light scattering, as we discussed earlier. However, if atoms and molecules are in very highly ordered arrays, as is the case in flat surface layers of solid matter (especially crystals and metals) or even in some amorphous materials, such as glasses, having only small-scale disorder, then the many Hertz dipoles can show constructive interference in certain directions. Accordingly, directional or specular reflection, just as occurs on silvered-glass mirrors, can be found at such solid surfaces. This gives a reflection at the same angle to the normal to the surface as the initial or incident radiation [6, 12]. If the reflection occurs due to refractive index differences (*Fresnel reflection*) then the amplitude of this reflectance depends on the refractive indices of the solid matter relative to the air (or if not air to the surrounding gas or vacuum). For the simplest case of normal incidence and non-absorbing media, the reflection is described by a simplified form of the *Fresnel equation*:

$$R = \left[\frac{(n_1 - n_2)}{(n_1 + n_2)}\right]^2 \tag{C4.1.5}$$

If the reflection occurs at a polished metal surface, then the resulting more-complete reflection can be predicted by Maxwell's equations, treating the metal as a conductive medium. If the metal is assumed to be a perfect electrical conductor, then the refection is theoretically predicted both by Maxwell's equations and by simple conservation of energy considerations to be 100%.

Refraction

Refraction describes the effect of a boundary between two optical materials on a light beam passing through it. Radiation incident at an angle, α, on the interface between media with different refractive indices (n_1 and n_2) is refracted at an angle, β, where, according to *Snell's law*:

$$n_1 \sin\alpha = n_2 \sin\beta \tag{C4.1.6}$$

Radiation, initially incident in a lower refractive index medium, is refracted towards the normal to the surface, so the angle of refraction becomes smaller, whereas, in the opposite direction it is refracted away from the normal.

Total internal reflection and evanescent field

Equation C4.1.6 suggests there is a critical angle at which the angle of refraction is 90°. If the angle of incidence increases further, then *total internal reflection* occurs, a behaviour only possible when light in an optically dense medium strikes an interface with a less dense medium. If the incident angle is greater than the critical angle, a simple explanation suggests radiation does not exit the denser medium, but is reflected back into this medium, as in figure C4.1.5. This effect has become very important in recent years for optical, or fibre-optical, waveguides (see Chapters A1.1 and A2.1).

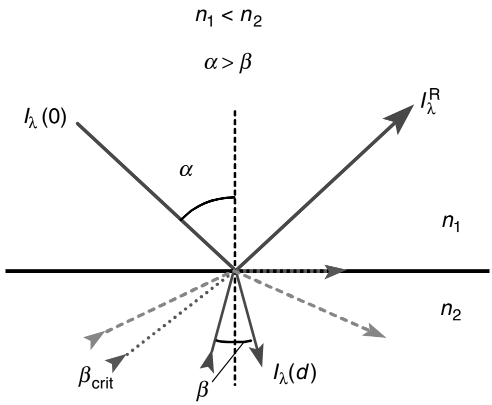


Figure C4.1.5. Behaviour of light incident on a dielectric interface, demonstrating refraction and total internal reflection.

We shall now move on to consider inelastic processes, where either light energy is lost totally, or where some light is re-emitted with a change in photon energy.

Absorption

Absorption is a loss of light in a material, involving an internal energy exchange, resulting in the destruction of a photon. It usually takes place only at specific wavelengths, corresponding to a defined transition between two energy states. Such a transition occurs when the natural frequency of charge displacement in the material is in resonance with the frequency of the incident radiation. The absorption involves a process that leads to a new *excited* (either stable or meta-stable) *energy state* in the material. In the case of electronic transitions in an atom, this involves creation of a new *electron density distribution*. In the case of molecules, it is possible instead to excite a new resonant, *vibrational* or *rotational,* mode of the molecule. An *electronic transition*, from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO), as described mathematically by equation C4.1.3, is demonstrated in figure C4.1.6. Note in figure C4.1.6, the phases of the orbitals are also given, in order to explain the symmetry of the orbitals, which causes specific transitions, shown by single, double or triple arrows. These demonstrate the differences in intensity of the transition. According to transition rules, only transitions between even and odd states are permitted.

These HOMO and LUMO states correspond to *bonding* or *antibonding orbitals* [10, 11]. The energy required for the transition is provided by the radiation and the process is known as (induced) absorption. Within this absorption band, anomalous dispersion is often observed (i.e. the refractive index increases with wavelength). Depending on the molecular environment and possible pathways of deactivation (see table C4.1.2), the new excited state can exist for a time varying over the wide range of 10^{-13}–10^{-3} s. From the *Schroedinger equation*, the corresponding energy states can be calculated as a set of *eigenvalues*, by using the electronic, vibrational, and rotational eigenfunctions and inserting the boundary conditions that are appropriate to the molecular structure. The relevant energy levels or states in the UV/VIS spectral region usually correspond to electronic levels, whereas in the infrared area they correspond to the energies of molecular vibrational modes.

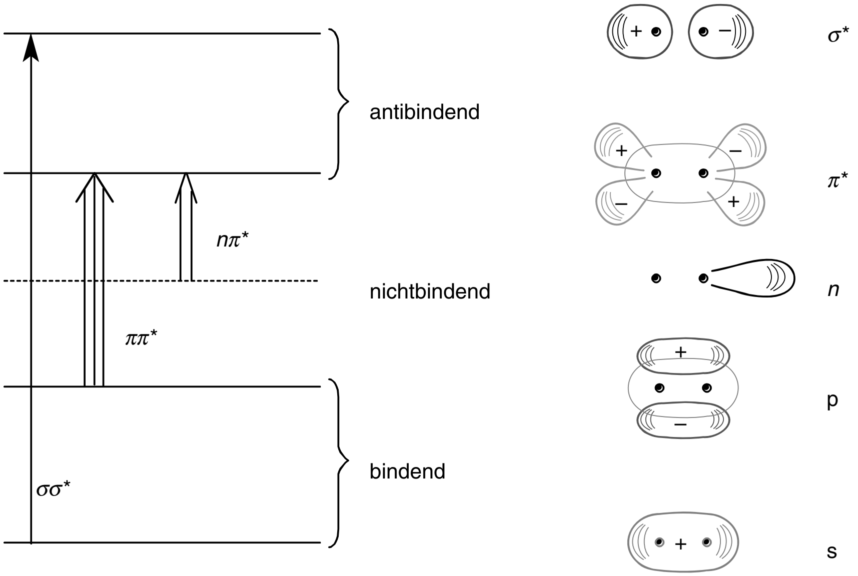


Figure C4.1.6. Bonding and anti-bonding orbitals (HOMO, LUMO) and their transition possibilities.

Energy level diagrams

In figure C4.1.7, a Jablonski energy level diagram for a typical organic molecule is depicted. It shows the energy levels and allowable transitions. It shows both electronic and vibrational states, but rotational states have been omitted to keep the diagram simple. At room temperature, most of the molecules are resting in the lowest electronic and vibrational state, so resonance absorption usually takes place to various excited higher-level vibrational levels or electronic states. To illustrate these, to the right of this energy level diagram, three types of resulting intensity spectra (absorbance (A), fluorescence (F), phosphorescence (P)) are shown as a function of wavenumber.

The relative strengths of absorptions depend on the momentum of the transitions and are determined by a set of spectroscopic selection rules. Radiative transitions requiring a change of electron spin are not allowed and, in organic molecules, a normal ground state has a singlet property. Thus, a normal absorption spectrum is for a transition from S_0 to S_1. Such a transition occurs within a very short time, typically $\sim 10^{-15}$ s. Usually only the first excited electronic state is important in spectroscopy, since all higher states have very short lifetimes (relaxation times). For organic molecules, usually paired spins are present (singlet states), whereas in inorganic transition metal complexes the ground state can be a triplet and other states with a large number of unpaired electrons may occur according to Hund's rule. An important example of a triplet ground state molecule is oxygen.

Electronic chromophores

Unlike the case for infrared spectra, where vibrational bands are very useful for the analysis of components, or of functional groups, the occupancy of, and energy between, the electronic levels (i.e. those which usually give rise to UV/VIS spectra) does not depend so much on the molecular structure or on interactions with the matrix. However, changing the electronic ground or excited states can lead to a

Table C4.1.2. Possible deactivation process from the vibrational ground state of the first excited singlet state.

Process type	Abbreviation	Name of process	Process description
Radiationless transition	Te	Thermal equilibration	Relaxation from a high vibrational level within the present electronic state. Lifetime depends on: Matrix Inner deactivation (transfer of energy in torsional vibrations)
	Ic	Internal conversion	Isoelectronic transition within the same energy level system from the vibrational ground state of a higher electronic state into the very high energy vibrational state of a lower electronic state
	Isc	Intersystem crossing, intercombination	Isoelectronic transition into another energy level system ($S \leftrightarrow T$), usually from the vibrational ground state in the electronically excited state; respective radiative transition is forbidden because of the spin inversion prohibition (except for heavy nuclei). Therefore phosphorescence is a 'forbidden' process
Spontaneous emission as a radiative process	F	Fluorescence	Without spin inversion from, e.g. S_1 to S_0 (provided that lifetime of the electronically excited state is 10^{-8} s) within singlet system
	P	Phosphorescence	Out of triplet into singlet system (provided that lifetime is 10^{-3} s) very low probability, only possible at low temperatures or in a matrix
Photochemical reactions			Photoinduced reaction starting from the S_1 term leading to ionization, cleavage, or in a bimolecular step to a new compound or an isomer (*trans–cis*), provided that lifetime in excited state is relatively long

shift in the relative position of the electronic level or to a change in the degree of polarization. In general, UV/VIS spectra are less easily used to characterize chemical components. However in some specific cases, for example steroids, incremental rules can be determined that allow both determination of, and discrimination between, some of these molecules.

Various transitions between electronic and vibrational levels are shown in the energy level diagram in figure C4.1.7. Depending on the electronic states involved, these transitions are either called $\sigma\sigma^*$, $\pi\pi^*$, or $n\pi^*$ transitions. They have been marked in figure C4.1.6. In aromatic compounds, electron-attracting or electron-repelling components, such as functional groups in the *ortho*-, *para*- or *meta*-positions of an aromatic ring, can affect these energy levels and they may also be changed by solvent effects or with pH changes.

Linewidth and line broadening effects

Although the basic absorption process occurs at discrete wavelengths, there are many mechanisms that can effectively broaden the effective absorption line.

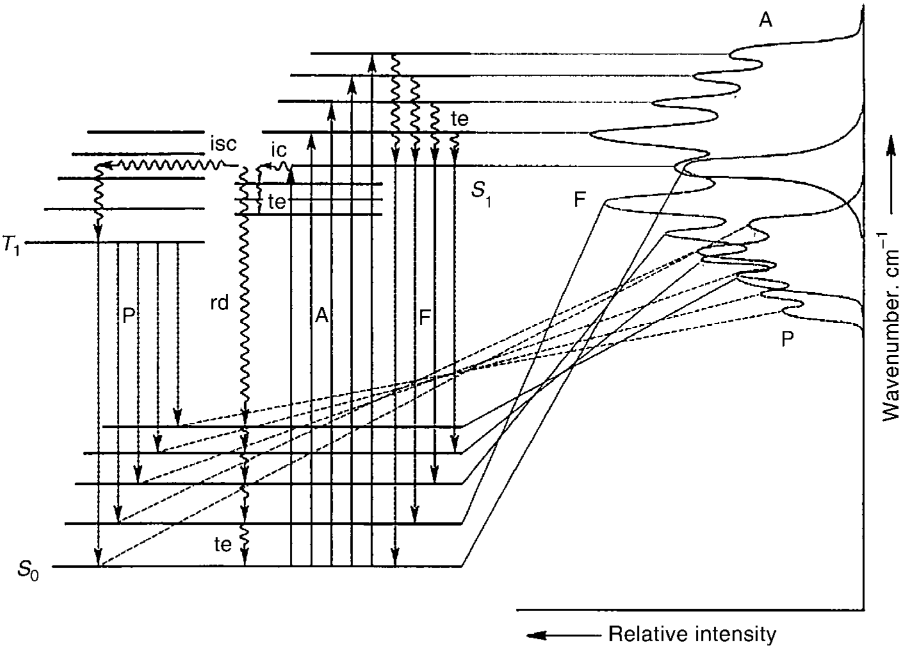


Figure C4.1.7. Jablonski energy level diagram, showing electronic and vibrational energy levels, and the resulting spectra for different transitions and pathways of deactivation.

The linewidth of absorptions is a very important aspect of spectroscopy [4], but we will present only a very brief discussion of the more important mechanisms. The main ways in which the effective linewidth of an absorption can be effected are:

- Velocity or movement effects, causing translational Doppler shifts
- Rotations of the molecule
- Interactions with other atoms in a molecule
- Interaction with other atoms or molecules by collisions
- Interaction with nearby atoms or molecules via electric field effects (Ligand fields)

Doppler shifts are usually only important for gas absorption lines. Single-direction wavelength shifts of this nature will always occur in any fluids that are moving rapidly (e.g. fast-moving gas jets), but the most usual effect is due to the rapid movement of gas molecules, which is defined by well-known gas laws, and of course depends strongly on the gas temperature.

Molecular rotations can occur according to the number of degrees of freedom of the molecule and these will cause a set of discrete energy (frequency) shifts, depending on the quantized rotational levels.

Interaction with other atoms in the molecule increases, for example, the number of degrees of freedom of vibrational and rotational transitions (e.g. number of possible modes of vibration) so molecules containing more than about six atoms usually tend to lack a fine-line absorption band structure.

A simple gas with atoms or molecules having a fine line absorption structure at low pressure will exhibit increasing broader lines as pressure is increased. This is due to an effect called *collision broadening* or *pressure broadening* and is due to the additional energy levels that arise due to collisions.

If the pressure is increased sufficiently, any closely spaced sets (manifolds) of narrow absorption lines will eventually merge into a continuous band.

Eventually, in the condensed state, the absorption is almost invariably of a broadband nature. This is because the large number of possible energy levels in absorbing bands can produce a near continuum. Electronic energy levels, for example, are affected by the electronic fields of nearby molecules, a phenomenon known as *Ligand field interaction.*

Bandshifts of electronic levels

Apart from line broadening, line-shifts can occur. As shown in figure C4.1.8, the $\pi\pi^*$ and $n\pi^*$ transitions are influenced differently by polar solvents. If a cyclohexane solvent is replaced by methanol, the polarity increases considerably and, accordingly, the $\pi\pi^*$ band is red-shifted by several nanometres, a so-called *bathochromic* effect, and the $n\pi^*$ band is shifted to the blue (*hypsochromic*). Such bandshifts can be used to obtain information regarding the properties of the energy levels and transitions. A change in intensity of the absorption band is described as either *hyperchromic* or *hypochromic*, respectively, depending on whether there is an increase or decrease in intensity.

Quantifying absorption levels in media

Absorption of samples is measured in units of *absorbance*, which is the log of the ratio of the light energy (I_0) entering a test specimen to the energy leaving (I_{out}).

The transmission, T, is given by I_0/I_{out} and the absorbance, A, is given by $\log_{10} 1/T$, so:

$$\text{Absorbance, } A = \log_{10}(I_{\mathrm{out}}/I_0)$$

An absorbance of 1 therefore implies only 10% transmission. One should take care in the use of the word ‘absorbance’, as defined above and as measured by instruments, as this word seems to suggest the optical loss is only due to absorption, whereas, in cases of turbid samples, it might actually arise from a combination of absorption and scattering.

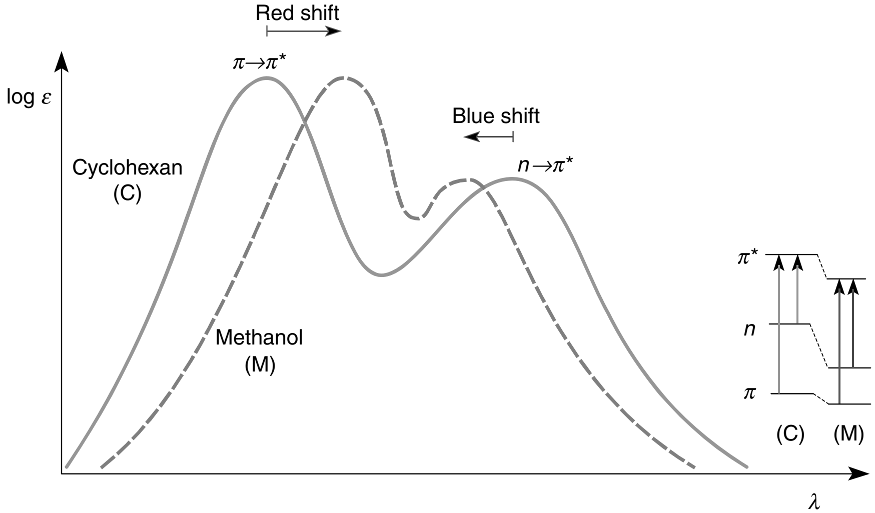


Figure C4.1.8. Illustration of band shifts caused by polar solvents.

The power, $P(\lambda)$, transmitted through a sample in a small wavelength interval at a centre wavelength λ, is given by Lambert's law:

$$P(\lambda) = P_0(\lambda)\exp[-\alpha(\lambda)\ell]$$

where $P_0(\lambda)$ is the power entering the sample in this wavelength interval, $\alpha(\lambda)$ is the attenuation coefficient of the material at wavelength λ, and ℓ is the optical path-length through the sample to the point at which $P(\lambda)$ is measured. This is only true for homogeneous non-turbid, non-fluorescent samples.

The sample can be said to have a transmission $T(\lambda)$, at the wavelength λ, where:

$$T(\lambda) = \exp[-\alpha(\lambda)\ell]$$

Alternatively, the sample can be said to have an absorbance $A(\lambda)$, where:

$$A(\lambda) = \log_{10}[1/T(\lambda)] = \log_{10}[P_0(\lambda)/P(\lambda)] = 0.43\alpha(\lambda)\ell$$

The factor 0.43, which is the numerical value of log(e), has to be included to account for the use of $\log_{10}$ for decadic absorbance calculations, whereas natural (to base e) exponents are normally used for attenuation coefficients.

According to the Beer–Lambert law, the value of A for an absorbing analyte is given by:

$$A = \log(I_0/I) = MCL$$

where M is the (decadic) molar extinction coefficient, C is the molar concentration and L is the optical path-length.

With merely measuring the intensity when using broadband radiation, there is a deviation from the above law whenever the absorption varies with wavelength. Then, each spectral component will obey the law, but the overall intensity will not. In all cases, when only measuring total intensity, care must be taken to avoid errors in calibration and/or absorption coefficient measurement. Without care, it is easy to get wrong values of absorption coefficients or to appear to be measuring the wrong compound. Fortunately, most transmission or absorption measurements are taken using spectrometers that show the full transmission or absorbance spectrum, so the behaviour of each component can be viewed separately. Even at single wavelengths, however, the law has the wrong dependence on concentration if there is clustering or association of molecules or if concentration-dependent or pH dependent chemical change occurs.

Frustrated total internal reflection, evanescent field absorption

In contrast to what is usually taught in classical optics, total internal reflection is not really 'total' [10] at the interface between the media. A fuller examination, using either classical geometric optics or quantum optics, predicts that part of the radiation will penetrate the interface and then, for a certain distance, be guided outside the optically denser medium, leading to a lateral shift in the apparent point of reflection from the interface. This field that extends into the less dense medium is called the *evanescent field.* In the medium having the lower refractive index, it decreases exponentially away from the interface.

Any absorbing medium within an evanescent-field penetration depth, d, of this field, can absorb the radiation and result in an attenuation of the otherwise total reflectance. The value of d (measured from the interface) is given by equation (C4.1.7):

$$d = \frac{\lambda}{2\pi\sqrt{n_2^2 \sin^2\Theta_2 - n_1^2}} \tag{C4.1.7}$$

This penetration depth is typically of order of half the wavelength of guided radiation. Thus absorption of the evanescent wave will reduce the intensity of the otherwise 'totally internally reflected' reflected light, via this electric field vector coupling.

This effect is called *frustrated total internal reflection*. It can be made practical use of for measurement of highly absorbing samples, which may absorb (or scatter) far too much light to be able to measure them by traditional methods. For example, complex samples such as paints, foodstuffs or biological materials, which might absorb or scatter light very strongly, can be analysed for their strongly absorbing components, by placing them in contact with a high index glass plate and measuring the internally reflected light.

Fluorescence

Fluorescence is an inelastic energy relaxation process, which can follow absorption of light in a material. In most materials, absorption of light (i.e. energy loss from a photon) merely causes heating of the absorbing material, with all the absorbed energy being converted to internal kinetic energy (for example, to excite molecular vibrations or phonons). However, in some materials, only part of the energy of an optically excited state is dissipated as heat and one or more photons, of lower energy than the incident one is radiated. This most commonly involves an internal process, which generates a photon and one or more phonons, with a total energy equal to that of the absorbed incident photon. The fluorescence is therefore usually at a longer wavelength (lower photon energy) than that of the incident light (this is called a *Stokes* process). The emission of light by fluorescence has no preferred direction and is said to be omni-directional. It is also randomly polarized. Of course a single photon fluorescent event must, by its nature, have the direction and polarization of the single photon emitted, but over a period of time the average scattered energy from many such events is omni-directional and so fluorescent light is randomly polarized. This aspect of fluorescence can be used, apart from wavelength and temporal-persistence (decay lifetime, after pulsed illumination) differences, to distinguish it from *Rayleigh* and/or *Raman* scattered light.

Fluorescence detection is a valuable technique in chemical sensing where it can be used to directly monitor certain fluorescent chemicals (e.g. poly-aromatic hydrocarbons, mineral oil, fluorescein tracer dye, etc). However, by deliberately introducing reactive *fluorophores*, which will act as a chemical indicator, it can also be used to monitor reactions or reagents having no fluorescence. Many optical techniques have been used [13] and many optical sensing applications of these are covered in detail in the proceedings of a series of 'Europt(r)ode' congresses. There is also an excellent textbook 'Optical Sensors', edited by R. Narayanaswamy and O.S. Wolfbeis [14].

Chemiluminescence and bioluminescence

Fluorescent light can arise as the result of chemical reactions, an effect known as *chemiluminescence*. The reaction must be of a type to leave electrons in excited states that can then decay radiatively, usually with the emission of visible photons. Such reactions are now very commonly seen in the plastic-tube-enclosed chemical 'light sticks' that children are often given, but which also have more serious uses for emergency lighting or as a distress signal at sea or in mountains. These lights operate by breaking chemical-filled ampules, enabling the reactants to mix and produce the light.

A compound commonly used to produce green light is called luminol, clearly deriving its name from its light-emitting properties. Luminol ($C_8H_7N_3O_2$) is known by several other chemical names, including 5-amino-2,3-dihydro-1,4-phthalazine-dione, *o*-aminophthalyl hydrazide, 3-aminophthathic hydrazide and *o*-aminophthaloyl hydrazide.

When luminol is added to a basic solution of oxidizing compounds, such as perborate, permanganate, hyperchlorite, iodine or hydrogen peroxide, in the presence of a metallic-ion catalyst, such as iron, manganese, copper, nickel or cobalt, it undergoes an oxidation reaction to produce excited electronic states, which decay to give green light. The strongest response is usually seen with hydrogen peroxide. Because photo-multipliers can detect single photon events in the darkened state, very low concentrations of oxidizing agents can be measured, including oxidizing gases such as ozone, chlorine and nitrogen dioxide. Numerous biochemicals can also cause a light-emitting reaction and hence be detected.

Nature has, however, used chemiluminescence long before man, involving a phenomenon known as *bioluminescence*, where the chemicals are generated within biological organisms. Well-known examples of this include the light from glowworms and fireflies and the dull glow arising from the bioluminescence of myriads of tiny organisms seen in agitated seawater, for example in the wake of boats. Many deep-sea creatures use bioluminescence, either to make themselves visible or to confuse prey. The reactions are essentially of the same type as that with luminol, except that the origin of the chemicals is from biological reactions.

Raman spectroscopy

Raman spectroscopy [15] relies on a form of inelastic scattering, where the periodic electric field of the incident radiation again induces a dipole moment in a material, but, in this case, there is an interaction with material vibrations, or phonons, which results in a change of photon energy. The effect depends on the polarizability, α, of the bond, at the equilibrium inter-nuclear distance and on the variation of this distance, Π, according to equation (C4.1.9) below. This equation has three terms—the first represents the elastic Rayleigh scattering (scattering at the same frequency as the incident radiation: 00 transition).

$$\alpha = \alpha_0 + \left(\frac{\partial \alpha}{\partial x}\right) x \tag{C4.1.8}$$

In the equation below, there are polarizability terms that have the effect of reducing or increasing the re-radiated frequency, compared with that of the incident light, by an amount equal to the molecular vibrational or/and rotational frequency. This causes so-called *Stokes* (downward) and *anti-Stokes* (upward) shifts in the re-radiated optical frequency. This light is referred to as Raman scattered light.

$$p = \underbrace{\alpha_0 E_0 \cos 2\pi \nu_0 t}_{\text{Rayleigh scattering}} + \frac{1}{2}\left(\frac{\partial \alpha}{\partial x}\right) x_0 E_0 [\cos 2\pi(\nu_0 - \nu_1)t + \cos 2\pi(\nu_0 + \nu_1)t] \tag{C4.1.9}$$

The coupling of vibrational eigenfrequencies (ν_1) of the scattering molecule to the exciting frequency (ν_0) is sometimes called a photon–phonon interaction process.

Raman spectroscopy occurs with selection rules that are different to those for normal vibrational-absorption infrared spectroscopy. Because of this, many users regard analysis by Raman scattering and by conventional absorption-based infrared spectroscopy to be complementary techniques. Unfortunately, the Raman scattering intensity is very weak, even when compared with Rayleigh scattering. Despite this, however, it is in many cases now a preferred technique, since it allows use of low-cost visible or near-infrared semiconductor lasers, with compact instrumentation.

Multi-photon spectroscopic processes

Apart from the simple processes described above, there are many light absorbing and/or re-emission processes that involve more than one photon. It is beyond the present text to discuss multi-photon spectroscopy, but it is worth making a few short comments.

Examples include 2-photon absorption, where it requires simultaneous involvement of two photons to excite an absorbing level (very unlikely occurrence, therefore a weak effect) or where the first photon excites to a meta-stable level and then the second photon excites it to a higher level from this intermediate level. Of course, many photons can be involved if an appropriate set of meta-stable levels is present.

C4.1.3 Spectroscopic methods and instrumentation

We shall now outline some of the methods and instrument types commonly used in spectroscopy. This has always been a very fast developing field, so only the basic methods will be discussed. The interested reader should consult specialist texts for more details of instrument design and manufacturers data for the latest performance figures. The fastest developing component areas are those of (a) low-noise detector arrays, both 1-D and 2-D types, and (b) compact high-power semiconductor laser sources for Raman and fluorescence excitation, where performances are now achieved that were barely conceivable only 10 years ago.

Just as an example of these, to illustrate the type of components needed, a compact fibre-compatible spectrometer source and detection unit with a broadband source and concave diffraction grating is shown in figure C4.1.9. In most cases, Xenon or halogen lamps are a convenient source of broadband white light, which (often with the help of a rear reflector, to aid launch efficiency) is coupled into a fibre optic guide and then directed to the measurement probe. Here, transmitted or reflected light is collected and, via the coupler guided back to a miniaturized spectrometer, which contains the diffraction grating and a photodiode diode array, such as a CCD detector chip.

Below, we shall discuss the components of typical spectrophotometer instruments, then go on to discuss instrumentation and measurement methods.

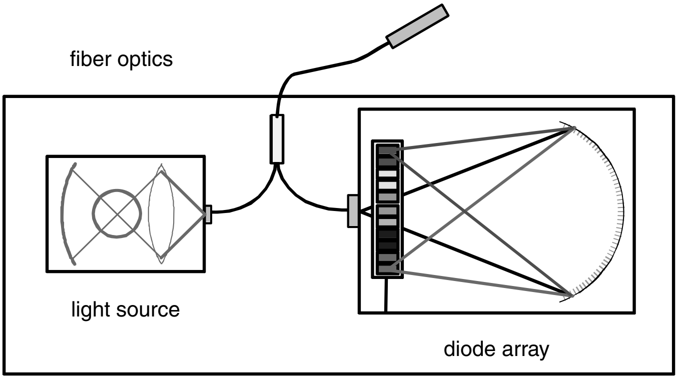


Figure C4.1.9. Spectrometer detection unit for reflectometric measurements, containing a white light source and a CCD camera. (A mirror is used in the light source to increase launch efficiency by imaging the filament on itself to ‘fill in’ gaps in the filament.)

C4.1.3.1 Spectrometer components

We shall now briefly review the components that are used in spectroscopic instruments, starting with light sources.

Light sources

Whenever an intense monochromatic light source or an intense light source that can be scanned over a small spectral range is required, the source of choice is nearly always a laser. This is clearly the optimum type of source for photon correlation spectroscopy or for Raman spectroscopy, as a monochromatic source is required. However, broadband sources, such as incandescent lamps, are desired for full-spectrum spectrophotometry, as intense narrowband sources, such as lasers or low-pressure mercury lamps, either cannot be used for observing spectral variations over such a wide range or would be prohibitively expensive.

Since most types of incandescent light sources were discussed in detail in Chapter A1.3, we shall cover only the most interesting spectral features of these broader band radiation sources, in particular types covering the important UV/VIS/NIR spectral regions. Tungsten lamps are low cost and stable sources, which give a useful output over the visible/near-infrared region (0.4–3.0 μm). Electrically heated thermal sources, such as Nernst emitters, or modern variants, are usually used to cover the infrared region beyond 3 μm.

High-pressure arc and discharge lamps are attractive sources in the visible and ultraviolet, as they can provide the desired broad spectral emission, but with significantly higher spectral intensity than tungsten lamps. In the visible and near IR region, the low cost and stable output of tungsten sources often means they are the preferred source. If operated at slightly reduced voltage, their lifetime is excellent too. However, discharge lamps are far more effective sources than tungsten in the blue and violet regions and are far better in the UV region. Thus, the type of discharge lamp most often used to cover these regions in commercial spectrophotometers is the Deuterium lamp, as this is the source that can provide the desired broadband performance at shorter optical wavelengths.

Figure C4.1.10 provides a simple comparison of discharge lamps.

It is worth mentioning one more broadband source that is very useful for spectroscopy, and that is the Xenon flashlamp. The output spectrum of this is, of course, very similar to that of the Xenon arc lamp shown above. However, because it is only operated in short pulses and at a very low duty cycle (maybe a 100 μS flash every 0.01 s), it can be driven at very high current density without destroying the electrodes. This provides an even brighter source than the continuously operated arc-lamp version, albeit for very short ‘on’ periods. These lamp sources also cover the UV, visible and part of the near infrared spectrum. They are not only very compact and have low average power consumption, but they can also provide a high peak energy output. This can give an optical signal at the detector that may be orders of magnitude higher than the thermal noise threshold of many optical receivers. They can therefore be used with optical systems, such as fluorimeters, where large energy losses often occur, yet still provide a good signal/noise ratio at the detector. Because of their pulsed output, they are also well suited for fluorescence lifetime measurements, provided the measured lifetimes are longer than the decay time of light output from the pulsed lamp.

Components for optical dispersion (wavelength separation and filtering)

Prisms

Many older spectrometers used prisms, which were constructed of optical glass or vitreous silica. These exhibit a change of refractive index with wavelength, thereby giving wavelength dependent refraction, as

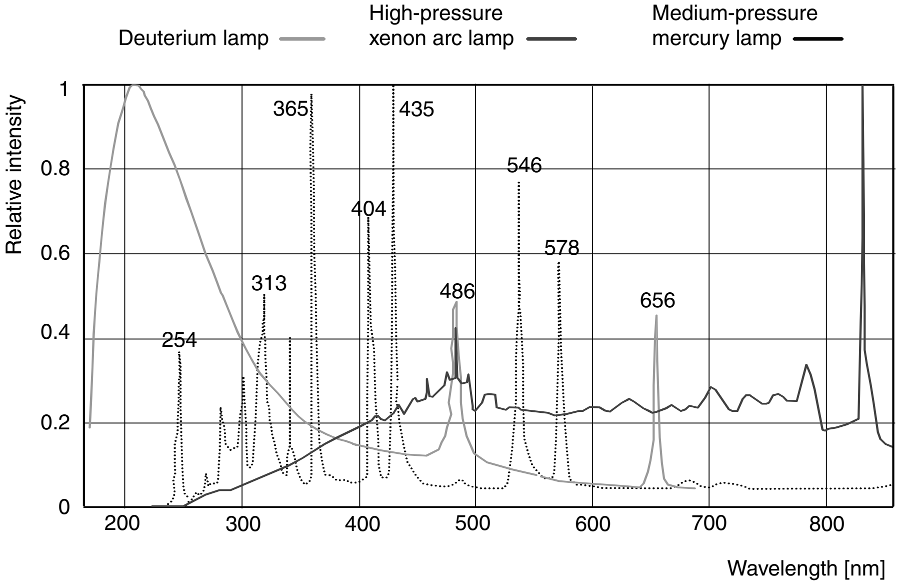


Figure C4.1.10. Comparison of emission spectra of a few discharge lamps and arc lamp sources.

shown in figure C4.1.11. Collimating the input light and refocusing onto a slit behind the prism, it is possible to separate out 'monochromatic' radiation.

In the ultraviolet, vitreous silica prisms are still occasionally used as wavelength dispersion elements, since they transmit well and their otherwise rather low material dispersion is a little higher in this region than in the visible, but even despite this, prisms are rarely used in spectrometers now. Because

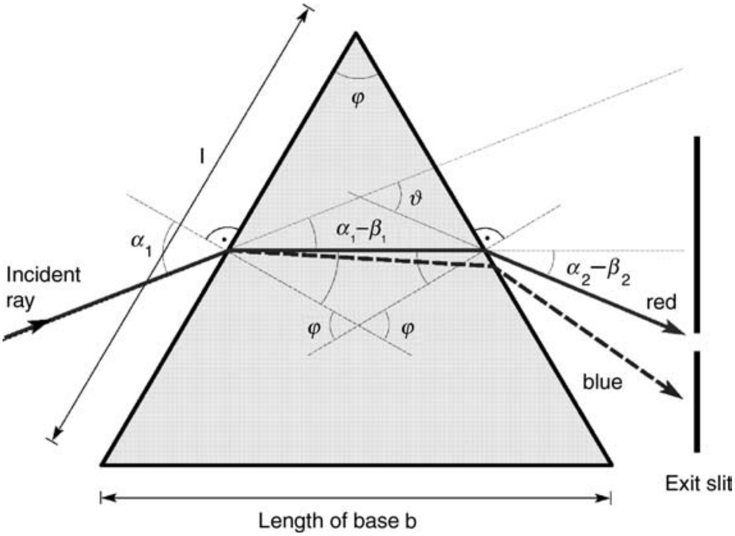


Figure C4.1.11. Dispersion using a prism: radiation incident on the prism is refracted and partially dispersed as it is deflected in φ direction towards the base of the prism. An exit slit selects the wavelength from the 'rainbow' of radiation. Usually, the red wavelength is less refracted than the blue one.

the prism material must be transparent and have a significant optical dispersion, greater problems with material selection arise in the mid-infrared, so this was the historical reason why diffraction gratings were first used in this wavelength region.

Diffraction gratings

Diffraction gratings are based on coherent interference of light after reflection from, or refraction through, an optical plate with multiple parallel grooves. The dispersion of a grating depends on the spacing of the grooves, according to the well-known Bragg condition that describes the *diffraction angle*, i.e. the angle at which there is constructive interference of the diffracted light. In the more common reflective diffraction gratings, the surface has parallel grooves, of a 'saw-tooth' cross-section, and these grooves have reflective surfaces. Apart from a high surface reflectivity, the angle of the 'saw-tooth' cross-section of the grooves to the plane of the surface is important to achieve good diffraction efficiency. The angle of the groove profile is called the blaze angle and this defines a blaze wavelength at which the diffracted power is maximized at the desired diffraction angle. The blaze angle can be chosen to maximize diffraction for the first order, or for a higher order if desired (see figure C4.1.12) [7].

Diffraction gratings were initially machined directly in a material that acted as the mirror reflector and the greater ease of making gratings that had a greater inter-groove spacing, meaning machining tolerances were less difficult, was another historical reason why gratings were first made for the IR region.

Most diffraction gratings are now made by one of two different processes. The first process, to make *'replica' gratings*, involves pressing (moulding) of an epoxy-resin-coated glass substrate against a ruled (or 'master') grating that has been previously coated with an aluminium or gold film, with pre-treatment to ensure this film will 'release' easily from the surface of the master when the epoxy has set and the two substrates are separated. This process duplicates the contours when the replica grating substrate, with the reflective metal film now attached to the outer surface of the resin, is pulled away from the master. The other production process, which is becoming ever more sophisticated, is to use photo-lithography to etch a grating in the material. The regular spacing is achieved by exposing photoresist using two converging parallel light beams, to give a regular set of parallel and equally spaced interference fringes. These gratings can now be made by exposing the resist and etching the substrate in small steps, in a way that allows control of the blaze angle. These are called *holographic gratings*, because of the interference, or holography process that is used to expose the photo-resist.

Optical detectors

Optical detectors and their associated amplifiers have been discussed in considerable detail in earlier sections, so there is no need to discuss them at length here. However, it is useful to comment that the most common single-element detectors used in wavelength-scanned spectrophotometers include photomultipliers (mainly for the UV range, below 400 nm), silicon photodiodes (visible and near infrared range from 400 to 1000 nm) and photoconductive detectors (usually to cover above 1000 nm wavelength range). Infrared detectors are often cooled, using thermo-electric Peltier devices or mechanical thermal-cycle engines, to improve their inherently worse noise performance.

Optical detector arrays

Detector arrays are most commonly used in the visible near IR region, primarily based on silicon (400–1050 nm) or GaInAs (500–1800 nm). One of the simplest approaches is to use a linear array of discrete photodiodes, each with its own separate pre-amplifier. It is more common in the silicon (400–1050 nm) wavelength range to use self-scanned arrays, however.

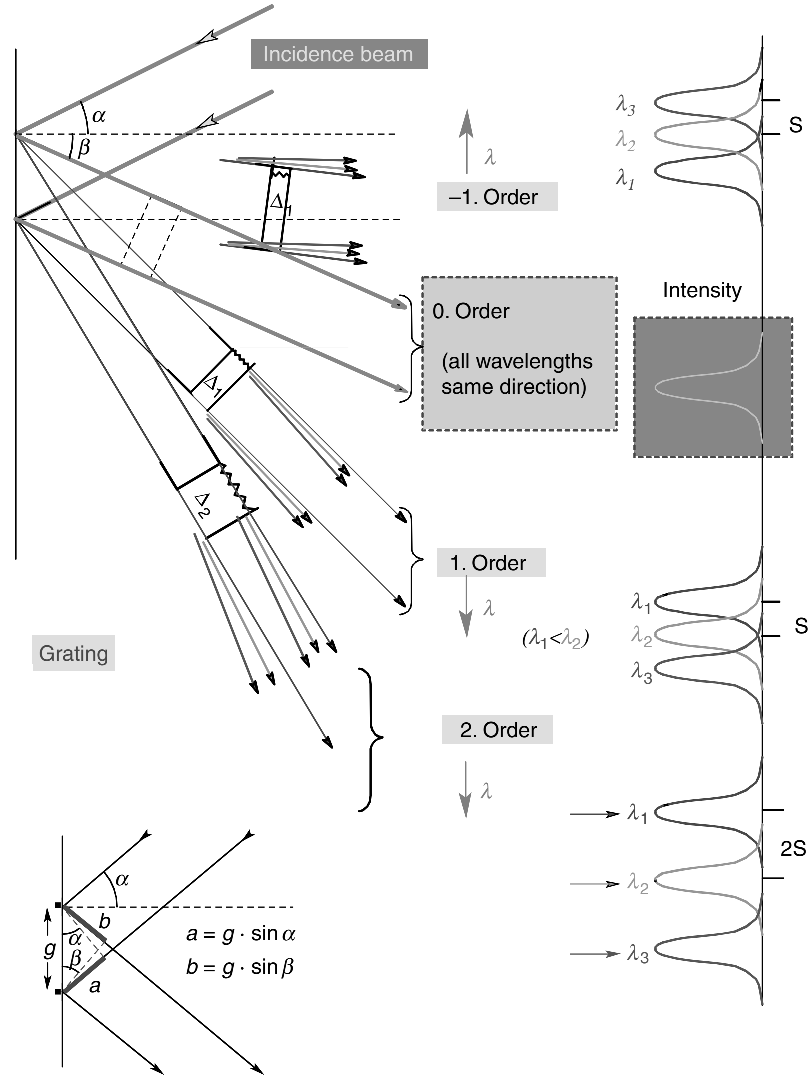

Incidence beam
α
β
Δ1
λ
–1. Order
λ3
λ2
λ1
S
Intensity
0. Order
(all wavelengths same direction)
Δ1
Δ2
1. Order
λ
(λ1<λ2)
λ1
λ2
λ3
S
Grating
2. Order
λ
λ1
λ2
λ3
2S
α
b
α
β
g
a
a = g · sin α
b = g · sin β

Figure C4.1.12. Dispersion at a diffraction grating: on the bottom left, the angle of incidence, α, corresponding to the strongest diffraction, depends on the groove spacing constant, g. The groove profile can be varied to give more light in a preferred direction and the grating is usually specified as having a blaze angle for a desired wavelength. In the upper part of the diagram, it can be seen that dispersion by the grating depends on the interference order. Higher orders suffer more dispersion, so hence provide better wavelength resolution.

Advances in silicon CCD and other forms of self-scanned detector arrays have been tremendous in recent years, initially driven by developments for camera technology, but with better performances being achieved for low-light-level cameras or for scientific uses (e.g. spectroscopy and astronomy). Individual pixel noise levels are now getting down to levels of below 0.01 photo-electrons per second in the most advanced cooled scientific versions. All of these types of low-noise self-scanned detectors effectively store photoelectrons (in the capacitor formed by the individual detector diode elements) until they are electronically scanned out in the read-out process. Note that detectors and detector arrays were reviewed extensively in earlier sections.

C4.1.3.2 Hardware instrumentation for spectroscopy

We shall now discuss the instrumental hardware that is used for some of the more-commonly used spectroscopic methods. Because of space limitations in this necessarily brief review, this covers only a selection of the basic methods that were introduced above.

Spectrometers: instruments for measuring optical intensity, as a function of wavelength

A spectrometer consists of a dispersive element, most commonly a diffraction grating, as described above, and a detection system. The latter is, most commonly, a silicon self-scanned CCD array (or other similar alternative), whenever visible or near infrared (400–1050 nm) measurements are taken.

The signal to noise of the spectral analysis and detection part of spectrometers is often enhanced by various opto-electronic means. One of the most common methods for visible/near-IR use is to employ multi-element self-scanned detector arrays with a fixed-grating light-dispersing system, as was shown in figure C4.1.9 above.

A large proportion of these low-noise self-scanned arrays are two-dimensional (2-D) types, with $n \times m$ detectors with up to several million elements, or *pixels*. Such 2-D arrays are clearly needed for cameras, either for low-light level terrestrial or for astronomical use, and they offer an attractive advantage for use in spectroscopy, allowing use of a technique called *spectral imaging*. If light that passes through an input slit of a spectrometer is dispersed by, for example, a diffraction grating, and then is focussed, or 'imaged', onto such a 2-D detector, light from each part of the input slit will produce its own individual spectrum or 'rainbow' image on this detector. This offers the possibility of measuring, independently, the spectrum of light that has entered at each point along the length of the input slit. The method therefore has the attractive property of measuring, simultaneously, the spectrum at many separate points on the input slit. There are many real-world applications of this. One, in the field of fibre-remoted spectroscopy, is to allow multiplexing of light exiting from many optical fibres held in the focal plane of the spectrometer. Each of these can be positioned in line, just as the single fibre in figure C4.1.9, but now, using a 2-D detector array in the output focal plane, a simultaneous measurement of the spectrum of light from each fibre can be made. This is then called an imaging spectrometer.

Other applications, without use of fibre optics, include imaging spectrometers for Earth-resources satellites, where a line image of a thin rectangular area on the surface of the Earth can be focussed onto the input slit and the spectrum of light from each point of the area can be independently analysed. As the satellite moves relative to the Earth, a full set of 2-D spectrally resolved images will therefore be monitored.

Alternatives to detector arrays

A common alternative to array technology, particularly in the mid- and far-IR region is to use some form of transform system, with a scanned optical filter with a complex multi-wavelength characteristic. Then, the detected signal is electronically decoded to recover the spectrum. Such methods will be discussed in more detail in a later section.

The spectrophotometer, an instrument to measure transmission or absorbance

Measurement of transmission loss usually involves the determination of light lost directly from a collimated light beam. This is normally performed in an instrument called a *spectrophotometer* and the result is usually required as a function of optical wavelength. The instrument normally includes a broadband light source, precise collimation optics to form the measurement beam or beams, some form of tuneable optical filter to select the wavelengths measured and finally a detector. This combination then allows transmission loss versus optical wavelength to be observed.

In the past, spectrophotometers were nearly always dual-beam instruments. In these, light from the source was split into two beams, which were alternately directed, firstly through a measurement sample path, where materials to be measured were placed and, secondly, via a reference optical path that contained no sample. This configuration allowed comparison (by signal division) of the transmission of each path, free of any intensity or spectral variations in the light source, or any spectral variations in the detector or in the collimation and filtering optics.

Many modern instruments are stable enough to first take a spectral measurement without a sample in the beam, store the result on a PC and then re-measure with the sample now inserted. Simple signal division, using the PC, then gives the transmission of the sample.

Indirect measurement of absorption using frustrated internal reflection methods

As mentioned above, frustrated internal reflection methods are useful for the measurement of highly absorbing samples, which may absorb (or scatter) far too much light to be able to measure them by traditional methods. For example, complex samples such as foodstuffs or biological materials, which might scatter light very strongly, can still be analysed to detect strongly absorbing species, by placing them in contact with a high index glass plate and measuring the internally reflected light. Such a plate, with suitable coupling optics, is a common inclusion in the sets of optional attachments that can be purchased for use with commercial spectrophotometers.

This evanescent field can be usefully used for various measurements, as demonstrated in figure C4.1.13.

Another application of this evanescent field is that fluorescent materials or *fluorophores* close to the interface can absorb this evanescent field radiation and induce fluorescence. The evanescent field can be used to monitor effects very close to the interface, since absorption clearly cannot take place beyond the penetration depth, so no fluorophores in the bulk are monitored.

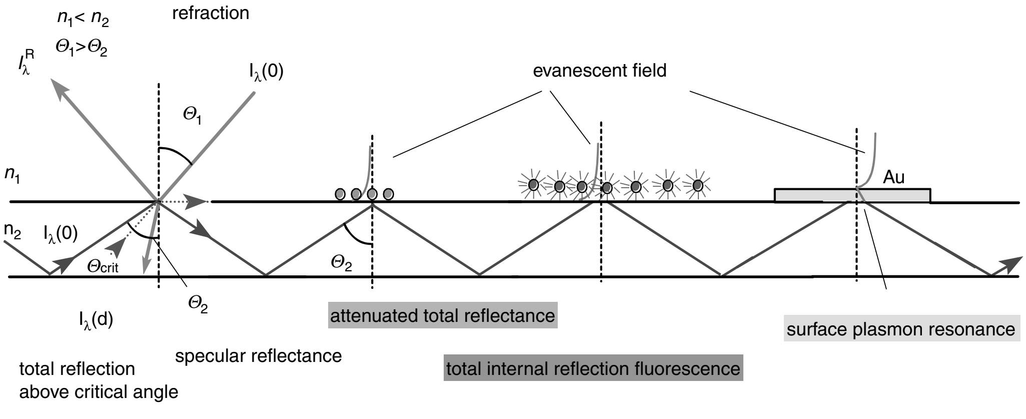


Figure C4.1.13. Illustration of the evanescent field caused by 'total internal reflection'. It can be partially absorbed (attenuated total reflection), or can excite fluorescence and surface plasmon resonance.

Measurement of indirect effects of absorption: photothermal and photoacoustic spectroscopy

Photothermal spectroscopy is a means of detecting or monitoring the absorption of materials, usually liquids or gases, by observing the resulting heat. The idea is that an intensity modulated intense light source is used to illuminate the sample and this raises its temperature when light energy is lost from the beam. An intense laser beam is most commonly used. This first heats the fluid and then becomes deflected by the transverse variation of refractive index that occurs when the resulting heated region induces convection currents. Such currents give rise to thermal gradients. In addition thermal-lensing effects can occur, which cause variation of beam diameter after passing through the fluid, because there is a higher temperature rise along the central axis of the beam.

Whichever method is used, the resulting deflections or beam diameter changes can be observed on optical detectors, often with some means of enhancing the sensitivity to changes in the beam position or diameter by using lenses and spatial filters.

Photoacoustic spectroscopy is a special case of photothermal spectroscopy, where the light from an intense source, again usually a laser, is intensity modulated, for example with a pulse, step or sinusoidal waveform, and the resulting changes in thermal expansion due to the heating are observed, usually with a traditional acoustic microphone. Gases have low specific heats and large expansion coefficients, so these are relatively easy to detect by this method. The light source can also be swept in optical frequency (or a broadband source can be passed through a swept optical filter) to allow spectral variations over an absorption line, or over a wider spectral region, to be observed.

A major advantage of the method for the measurement of liquids is that it can be used with turbid samples, for example ones containing small bubbles, or a suspension of non-absorbing scattering particles. In a conventional spectrometer, it is clearly not normally possible to distinguish between the light lost from the collimated beam due to elastic scattering (turbidity) and the light lost due to absorption.

Monitoring of chemiluminescence

As discussed earlier, when luminol is added to a basic solution of oxidizing compounds, such as perborate, permanganate, hyperchlorite, iodine or hydrogen peroxide, in the presence of a metallic-ion catalyst, such as iron, manganese, copper, nickel or cobalt, it undergoes an oxidation reaction to produce the excited states that decay to give green light. The strongest response is usually seen with hydrogen peroxide. In order to detect, and measure, this weak green light, it is best to use a photomultiplier, with some form of light-reflecting cavity (light-integrating chamber) to ensure most of the light hits the sensitive photocathode. Because photo-multipliers have good green-light sensitivity and can detect single photon events in the darkened state, very low concentrations of oxidizing agents can be measured, including hazardous oxidizing gases such as ozone, chlorine and nitrogen dioxide. Numerous biochemicals can also cause a light-emitting reaction and hence be detected. A particularly useful reaction for law enforcement is the one that luminol has with blood, enabling crime scenes to be sprayed with this compound and then be viewed in the dark, when telltale glows appear, wherever traces of blood are present. Chemoluminescence is the basis of a number of commercial chemical sensors for important biochemicals.

Instrumentation for fluorescence spectroscopy

Fluorescence spectroscopy involves illumination of a sample with a monochromatic or filtered light source and observing the re-radiated signal, which is almost invariably at a longer wavelength than the incident light. It is common to observe either the fluorescence spectrum or the fluorescent decay curve, following pulsed excitation from a pulsed source. The latter is usually a pulsed laser or filtered light from a Xenon flashlamp.

For fluorescence spectra, the most common method is to use a modified spectrophotometer where part of the omni-directional fluorescent light is observed instead of the transmitted light. This can be done either with a dedicated instrument, or can be performed using a standard commercial spectrophotometer instrument fitted with a fluorescence attachment, which has appropriate optics to gather fluorescent light. There is no need to modulate the light source for this measurement, unless this is needed to allow lock-in (phase-sensitive) detection methods to remove signals from unmodulated background light. When performing these measurements, considerable care has to be taken with the following filter-related aspects:

- Removal of any longer-wavelength 'side-band' light from the light source, which could be elastically scattered in the sample or the instrument and mistaken for fluorescence.
- Very careful filtering of the 'fluorescent' light signal to remove any elastically scattered source light.

The first aspect of source light filtering is less of a problem when using laser sources, although semiconductor lasers can have some residual spontaneous-emission component at longer wavelengths and gas or ion lasers can have residual light from the plasma discharge. Particular care is needed with incandescent, arc lamp or Xenon flashlamp sources, with their broadband emission. In a commercial spectrophotometer instrument, or a dedicated instrument (spectrofluorimeter), the built-in grating spectrometer acts as a very good filter, provided 'stray light' levels are low and also has the advantage that the excitation wavelength can be tuned if desired. Additional rejection of long wavelength light is usually done with 'short-pass' dichroic multi-layer edge-filters. The problem of removing elastically-scattered source light from the 'fluorescent' light signal can be done in several ways. Narrow-band dielectric laser mirrors make excellent rejection filters in transmission mode, as these can be designed with a reflectivity of 99.9% or higher. Dichroic long-pass edge filters are also now available with excellent performance. In addition, there is a wide range of long-pass optical glass filters, which have semiconductor-band-edge type transmission behaviour, commonly having short-wavelength (i.e. shorter than the band-edge) absorbance as high as 106 in a 2-mm-thick filter. Care must be taken with these, however, as many of these fluoresce at longer wavelengths, so the first filtering stage is preferably done with a dielectric filter. Note that these filtering problems are even more acute for Raman scattering, so will be discussed further in the section below.

When it is desired to examine fluorescent lifetimes, the severity of the filtering problem is reduced by several orders of magnitude as the pulsed source will usually be no longer emitting as the intensity-decay curve of the fluorescence it initiated is now observed. However, strong light pulses can upset many sensitive detection systems or associated electronics, so even here some degree of filtering is still desirable to remove light at the incident wavelength and some source filtering may be necessary too if there is a spontaneous light decay in the laser source.

When measuring fluorescent lifetime, a fast detector may be needed. Fluorescent decay is commonly in the form of an exponentially-decaying curve, but lifetimes can typically vary from days, in the case of phosphorescence, to less than nanoseconds. (Important examples of samples having decay times in nanoseconds are the organic dyes often used in dye laser systems and some semiconductor samples with short excited state lifetimes.) When measuring very fast fluorescent decays, it is common to use photon-counting systems using photomultiplier (PMT) detectors. These have the advantage of high internal gain of the initial photo-electrons, so the input noise level of even a fast electronic pre-amplifier is easily exceeded. Also, as the desired detection time is reduced, by using a fast PMT and amplifier, the effective peak current represented by a given-size 'bunch' of electrons, which will arrive at the anode from a single-photon event, actually increases as the time interval over which it is measured reduces (Current = charge/time). Thus, using fast photon-counting technology, where each photon count is fed into sets of digital registers according to its time of arrival, very fast fluorescent decay curves can be measured. It is now becoming more common to design photon counting systems with avalanche photodiode detectors, which are operated in a so-called 'Geiger' mode, where the incoming photon causes a full (but reversible, after the bias voltage is re-applied) conductive 'breakdown' of the reverse-biased detector diode.

If there is more than one fluorophore or more than one fluorescent decay process occurring, the decay make take the form of a bi- or multi-exponential decay curve, equivalent to linear addition of two or more exponentially-decaying functions. In simple cases, these can be separated with software, but in some cases appropriate choice of excitation wavelength may help to isolate individual curves in mixtures of fluorophores.

Another common way of measuring fluorescence lifetime is to intensity-modulate the source with a periodic waveform, usually a sinusoidal or square-wave signal, and observe the phase delay in the fluorescent signal intensity waveform, relative to that of the incident signal. This is commonly done using an electronic system, which can multiply the detected fluorescence signal by an electronic signal analogue of the incident-intensity waveform, and then averages the product over a set time interval. If desired, more than one multiplication can be performed, using reference signals with different phases. Such an electronic detector system has several names, which include vector voltmeter, lock-in amplifier or phase-sensitive detector. Essentially it enables the phase difference between the two signals to be measured. The advantage is that cheaper, simpler detectors, such as silicon photodiodes, can now be used, as the illumination duty cycle is improved (to 50%, in the case of square-wave modulation), which helps to improve the signal to noise ratio, but the disadvantage is that the shape of the decay curve cannot be seen. Another significant disadvantage is that the system now requires much better optical filtering, as any residual optical crosstalk, where elastically scattered light from the source might be detected, will alter the effective phase of the detected 'fluorescent' signal.

An important feature is that fluorescence detection can be performed with highly scattering samples, such as roughly cut or even powdered materials, and can be used to analyse the surface of opaque materials, as a clear transparent sample is not required. Also as the transmitted light level is not measured, very little sample preparation is needed. Another minor advantage is the slightly shorter excitation wavelength means it can be focussed to a slightly smaller diffraction-limited spot, enabling its use in fluorescence microscopes, which excite the specimen via reasonably conventional optics. These advantages apply even more to Raman spectroscopy, which will be dealt with below, so they will be repeated again there for greater emphasis.

A particular problem with fluorescence detection is that many materials will fluoresce, particularly if illuminated with UV light. These include, starting with particularly troublesome examples, organic dyes, compounds of rare-earth metals, chlorophyll, ruby, long-pass optical absorption filters, mineral oils, human sweat, many adhesives and even some optical glasses.

Instrumentation for Raman spectroscopy

Raman spectroscopy can use conventional optical materials with low-cost semiconductor lasers and high-performance cooled CCD detector arrays. A major advantage is that it can be used with highly-scattering samples, such as roughly-cut or even powdered materials, and even analyse the surface of opaque materials. Obviously, a clear transparent sample is not required, as the transmitted light level is not measured, so very little sample preparation is needed. Another major advantage is the short excitation wavelength that means it can be focussed to a smaller diffraction-limited spot, enabling its use in Raman microscopes, which excite the specimen via reasonably conventional optics, and allow spatial resolution that would be impossible with direct IR excitation. Confocal microscopy is also possible using Raman methods.

Raman spectroscopy requires a laser source to excite the very weak Raman scattering and a highly sensitive spectrometer to detect the very weak scattered light. The sample is illuminated with focussed laser light and the Raman scattered light is collected. Its power spectrum is analysed, as a function of either optical wavelength, and often presented in terms of its power spectrum in wavenumbers, an optical unit of frequency more commonly used by chemists. Even in clear samples, such as optical glasses, which are only very weak Rayleigh scatterers, the Raman light intensity is usually several orders of magnitude below that even from this elastic scattering. In early days of Raman work, it was necessary to use either laser sources having a power of 1 Watt or more, with very sensitive photomultiplier detection systems, having substantial post-detector averaging to measure the low levels of signal.

A major part of the problem was that highly-scattering samples, such as cut materials or powders, can have elastic scattering levels 5 or more orders of magnitude higher than clear samples. Thus, it is necessary to separate out the very much stronger elastic scattering that can occur at the same wavelength as that of the incident laser light, and which might be between 5 and 9 orders of magnitude higher. It was the extensive optical filtering and the scanned (one-wavelength-at-a-time) nature of the spectrometers used that caused the systems to have very poor optical efficiency.

The wavelength filtering to achieve separation usually requires at least two or maybe three stages of optical filtering to recover the desired Raman light, free from the undesired elastically-scattered light. In the early days of Raman instrumentation, it was common practice to use rather large *double* or *triple monochromators* (i.e. 2 or 3 cascaded grating spectrometers), to reduce crosstalk from so-called *stray light* in the spectrometers themselves, which might otherwise have caused undesirable elastically scattered light to strike the detector array sites on which the desired Raman light would normally be diffracted. Even then, simply using multiple monochromators was not always sufficient to achieve full separation of elastically scattered light, particularly as any scattered incident light could cause fluorescence on any contaminated optical surfaces it struck. If this occurred, no degree of spectral selection could eliminate it, as the fluorescence band could then overlap the desired Raman bands.

Second generation Raman systems reduced this problem by directing the light though a holographic *Raman-notch filter*, which is a compact optical filter designed just to reject a very narrow band of light, centred at the incident laser wavelength. It was easier to construct this compact component using low-fluorescence materials, thereby easing the problem of further separation in the dispersive spectrometers. With such a filter present in the detection system, a single spectrometer could be used, saving both space and cost. However, at the same time, another modern technology emerged in the form of greatly-improved self-scanned silicon detector arrays, having very low noise and high quantum efficiency and being capable of providing even better low-light performance than photomulipliers on each tiny pixel in the array. This has allowed a compact state-of-the-art detector array to be placed in the output focal plane of a single monochromator, eliminating the need for a narrow output slit that rejected most of the light and now allowing all the wavelength components of the Raman spectrum wavelengths to be measured at the same time.

A more recent alternative to the Raman notch filter is the use of dichroic band-edge filters. These interference filters, made in the usual way with a stack of dielectric layers on glass, have a long-wavelength pass characteristic that removes not only the undesirable elastically-scattered light, but also the anti-Stokes Raman light. These new filters are highly effective, are of relatively low cost, but are far more stable than the holographic ones used for notch filters, which were based on various polymer materials. A minor disadvantage is the removal of anti-Stokes signals, but as these are not used in 90% of applications, this is not a major disadvantage.

The most troublesome residual problem with Raman systems is, as it always has been, that of undesirable background signals due to fluorescence. It has been described how the effects of this could be reduced in the spectrometer, but sample fluorescence is more difficult to remove. Fortunately, using near-infrared illumination, typically at 800 nm wavelength or above, goes a long way towards reducing the background fluorescence of most common materials and contaminants. There are also various signal-subtraction methods that can be used. These can, for example, take advantage of the polarization dependence of Raman or of the temporal decay behaviour of fluorescence, but these are beyond the scope of this introductory text.

All the above developments have given many orders of magnitude improvement in the spectrometer and detection system, to a situation where now, Raman scattering is often a preferred technique to absorption spectroscopy. To repeat the Raman-specific advantages mentioned above, a major feature is that, as a good optical transmission is not required, the method can be used with highly-scattering samples such as roughly-cut or powdered materials, and even with opaque ones. It can also be used for high-resolution microscopy, because of the smaller diffraction limited spot at the shorter excitation wavelengths.

It is beyond the scope of this text to go into detail on other techniques for enhancing Raman signals, but it is useful to mention two briefly. The first is resonance Raman scattering, where the excitation wavelength is close to an absorption line of the material to be monitored. The second is surface-enhanced Raman spectroscopy (SERS), where the use of a metal surface, or a surface covered with metal particles, can enhance, greatly, the signal from Raman scattering. A few materials, such as silver or silver particles, are highly effective for this. Both of the above methods can enhance the Raman signal by between 4 and 8 orders of magnitude, depending on conditions. Being a surface effect, however, the SERS method is clearly very sensitive to surface preparation conditions and to subsequent treatment or contamination.

Instrumentation for photon correlation spectroscopy

The use of photon correlation spectroscopy for particle detection was discussed earlier. It is particularly useful for determination of particle size, over a range of a few nm to a few μm, simply by measurements of scattered signals. Very small particles will undergo fast random movements, as molecular collisions move them about (Brownian motion). In a conventional instrument, a polarized monochromatic TEM_{00} laser beam is usually used to illuminate the sample by focussing to a narrow beam waist, and a fixed detector is used to observe the scattered light from the beam-waist area. The laser source must be stable in intensity to a small fraction of 1% and usually has a power of a few mW. Originally gas lasers were used, but now compact semiconductor lasers are replacing them.

A single spatial mode (single ‘speckle’) in the far field will exhibit more rapid intensity changes with small scattering particles present when compared with the same conditions for larger particles. In order to detect the changes of phase, the scattered light is traditionally imaged through a very small hole to act as a spatial filter, in order to provide the greatest intensity modulation index (greatest fractional change in optical intensity, as the optical phase changes). Clearly, a large detector, without any pinhole aperture is unsuitable, as it would average the light from many interference points, so would see a smaller modulation index.

For successful operation, all the optics must be very clean and additional opaque baffles are often used to reduce stray light. Samples must be prepared very carefully to avoid highly-scattering large dust particles or bubbles and clearly particle aggregation (clustering) must be avoided.

Following optical detection of the intensity changes, electrical spectral analysis (frequency analysis) of the signal scattered, for example at 90°, could potentially yield valuable information on particle size, with smaller particles giving higher intensity modulation frequencies. Unfortunately, very small particles also scatter very weakly and, to act as an effective spatial filter, the receiving aperture has to be very small. As a result, the received photon flux is usually very low, sometimes only a few photons per second, making frequency analysis of a very noisy (random photon event) signal much more difficult. A preferred alternative method, which is more suitable for use at low photon flux levels, is to use a method called photon correlation spectroscopy (PCS). Here, instead of analysing the frequency of detected signals, the times of arrival of many individual photon pulses are correlated. The decay time of what, for monodisperse (single-size) particles, is usually an exponentially-decaying correlation function, can be derived using digital correlator systems. As stated earlier, this can, with knowledge of the temperature and viscosity of the fluid in which the particles are suspended, be related to the *hydrodynamic radius* of the particles using the Stokes–Einstein equation. The method works best with single-size (monodisperse) particles, but more complex correlation functions from suspensions of two or even three particle types/sizes can be inverted using the Laplace transformation.

The photon correlation method works well at these very low light flux levels, but requires the use of single photon counting detectors, such as photomultipliers or silicon avalanche photodiodes (APDs). Significant developments have been made to actively quench APD photon counters that are operated in *avalanche mode*, at very high reverse bias, to allow fast recovery from their photon-induced breakdown and be ready to detect a subsequent photon. This allows instrumentation to take advantage of their superior quantum efficiency in the near infrared [16].

Other developments have been on the design of fast correlators to process the photon count signals and recover particle size information. PCS can typically achieve an accuracy of order 0.5% in particle size on monodisperse samples. With more sophisticated signal processing, it is possible, provided conditions are suitable, to derive estimates of particle size distribution, polydispersity, molecular weight estimates (using Svedberg's equation), rotational diffusion behaviour and particle shape and many other parameters. The greatest practical problem is usually when large particles are present, either as an undesirable contaminant or as an inevitable feature of the sample to be monitored, as scattering from just a few large particles can often be intense enough to totally dominate the weak signals from very many much smaller ones.

Measurement with optical fibre optic leads

We shall now discuss how optical fibres can be used with various forms of spectroscopic instrumentation, and discuss the advantages and penalties of using them. Generally, if they can be used efficiently, optical fibre leads offer tremendous advantages. First, expensive instrumentation can stay in a safe laboratory environment, free from risk of damage from chemicals, weather, or careless handling. Second, the remote measurement probe can be very small, robust and immune to chemical attack. Third, there is no need to transport chemicals or other samples to the instrument, so real-time on-line measurements are possible, with no risk to personnel.

In-fibre light delivery and collection for transmission measurements

Transmission (and hence absorption or turbidity) measurements can be most easily achieved over optical fibre paths by using a commercial spectrophotometer with specially-designed extension leads. Many manufacturers now offer such systems as standard attachments. Typically, they have a unit that fits into the cell compartment of a standard instrument, with a first (focusing) lens that takes the collimated light that would normally pass through the sample chamber and focusses it instead into a large-core-diameter (usually $> 200\ \mu\text{m}$) optical fibre down-lead and with a second lens that re-collimates light, that has returned back from the sample area (via the return fibre lead), to reform a low-divergence beam suitable for passage back into the detection section of the instrument.

There is also a remote measurement cell in the sample-probe, connected to the remote end of both these fibre leads. In this remote sample area, a first lens collimates light (coming from the spectrometer, via the down-lead) into a local interrogation beam. This beam passes through the remote measurement cell, after which a second lens collects the light and refocuses it into the fibre return lead going to the spectrometer instrument. Such optical transformations lead to inevitable losses of optical power (due to reflections, aberrations and misalignments) of typically 10–20 dB (equivalent to losing 1–2 units of absorbance in the measurement range). However, most modern spectrophotometer instruments have a typical dynamic range of >50 dB, so this optical loss is a price that many users are prepared to pay in order to achieve a useful remote measurement capability.

It should be noted that the optical power losses usually occur mainly due to misaligments and the imperfections of the focusing and re-collimation optics, plus Fresnel reflection losses at interfaces, rather than arising from fibre transmission losses. If suitably collimated beams were to be available in the instrument, if large core diameter fibres could be used to connect to and from the probe and if all optics, including fibre ends, could be anti-reflection coated, there should really be very little loss penalty. Such losses therefore arise primarily because of the need for the fibre leads to be as flexible as possible (so hence choice of small diameter fibres) and the usual need to compromize design on grounds of cost (although, like all such attachments, they are very expensive to buy.)

There are many other probe head designs that are possible. The simplest design, for use with measurement samples showing very strong absorption, is simply to have a probe that holds the ends of the down-lead and return fibre in axial alignment, facing each other across a small measurement gap, where the sample is then allowed to enter. Losses are low for fibre end spacing of the same order as the fibre diameter or less, but rapidly increase with larger gaps. The probe is far easier to miniaturize and to handle if the fibre down-lead and return lead are sheathed, in parallel alignment, in one cable. Use of such a cable can be accommodated using a right-angled prism or other retro-reflecting device to deflect the beam in the probe tip through the desired 180° that allows it to first leave the outgoing fibre, pass through a sample and then enter the return fibre. Use of a directional fibre coupler at the instrument end allows use of a single fibre, but then any residual retro-reflection from the fibre end will be present as a crosstalk signal, adding light signal components that have not passed through the medium.

Clearly there are many variants of such optical probes, some involving more complex optics (e.g. multi-pass probes), some constructed from more exotic materials to withstand corrosive chemicals. A very simple option that has often been used with such single fibre probes, for monitoring the transmission of chemical indicators, is to dissolve the indicator in a polymer that is permeable to the chemical to be detected and also incorporate strongly scattering particles in the polymer. When a small piece of such a polymer is formed on the fibre end, the particles give rise to strong backscattered light and the return fibre guides a portion of this to the detection system. This backscatter light had of course to pass through the indicator polymer in its path to and from each scattering particle, so the returning light is subject to spectral filtering by the indicator. Although this is a very lossy arrangement, it is extremely cheap and simple and has formed the basis of many chemical sensors, for example ones using pH indicators.

In-fibre light delivery and collection for Raman scattering and fluorescence

Raman scattering and fluorescence can also be measured via fibre leads, but the use of fibres causes far more loss of light due to the wide-angle re-radiation patterns characteristic of both of these. However, the potential value of these methods, particularly of Raman scattering, for chemical sensing has meant workers will continue to persevere to get useful performance, despite the low return light levels encountered with fibre-coupled systems. Both these mechanisms involve excitation of a sample with light, usually at a wavelength shorter than the scattered light to be observed, and then the re-emitted light is collected and narrow-band filtered.

The loss due to launching of the excitation laser light into a fibre is usually negligible with Raman, as narrow-line laser sources are used, but ultimately the launch power limit may be set by non-linear processes or, in the case of large-core multimode fibres, by optical damage thresholds. Similar excitation can be used for low-level fluorescence monitoring, provided no photo-bleaching or other photo-degradation of the monitored substance can occur at high illumination intensity. The main potential loss is therefore that of light collection. Minimum additional loss due to collection of Raman light via optical fibres is at least a factor of 90 worse than using conventional optics, thus making the already low light levels about 2 orders of magnitude worse. Despite this, a number of fibre-remoted Raman chemical sensor probes are appearing as commercial items.

Evanescent field methods of in-fibre light delivery and collection: frustrated total internal reflection

Frustrated internal reflection measurements are clearly attractive for use with multimode optical fibres, as the guidance in the fibre depends on such internal reflections. As stated already, total internal reflectance is often not total. One marked departure from traditional school teaching occurs during reflection at curved interfaces, where light can, under certain circumstances, be lost by radiation into the less dense medium, even when the angle of incidence is above the well-known critical angle. This occurs in multimode optical fibres if light is launched to excite skew rays, provided the light is launched at an angle to the fibre axis that is too large to expect guidance by total internal reflection, assuming rays were passing through the fibre axis. Thus, even if the actual angle of incidence on the fibre core cladding interface for these skew rays is greater than the calculated critical angle, light will be lost into the cladding to form what are called *leaky rays*.

However, even with light incident on flat surfaces, so assuming no leaky rays are possible, there is always a small penetration of light into the less dense medium. Another aspect that is not normally taught in elementary optical texts is that the reflected light beam, although being reflected at the same angle to the normal as the incident light, undergoes a lateral shift in the direction of the interface plane between the two media. It is said to suffer a so-called *Goos-Hänchen* shift (this is like a displacement of the ‘point’ of reflection at the interface). A fuller examination, using either classical geometric optics or quantum optics, predicts that part of the radiation, the *evanescent field*, penetrates the interface and then, for a certain distance, is guided outside the optically denser medium, leading, when it returns, to a lateral shift in the apparent point of reflection from the interface.

The concept of the *evanescent field*, which decreases exponentially away from the interface in the medium having the lower refractive index, was introduced earlier. This evanescent field can be usefully used for various fibre-based experiments, as shown in figure C4.1.13.

As stated earlier, material or molecules within a penetration depth, d of this field can absorb the radiation and result in an attenuated total reflectance. The absorption can be enhanced further using thin metal layers on the fibre, to cause evanescent field enhancement due to plasmon resonance. This effect will be discussed in more detail later. Another application of this evanescent field is that fluorescent materials or *fluorophores* close to the interface can absorb this evanescent field radiation and induce fluorescence. The evanescent field can be used to monitor effects very close to the interface, since absorption clearly cannot take place beyond the penetration depth, so no fluorophores in the bulk are monitored.

In-fibre light delivery and collection for photon correlation spectroscopy (PCS)

In photon correlation spectroscopy systems, monomode fibres are very well suited for both delivery and collection of light. A single-mode optical fibre not only makes an excellent delivery medium for a beam launched efficiently from a gas or semiconductor laser, but it can also form a near-ideal single-spatial-mode optical filter to replace the conventional pinhole used for traditional systems. The PCS technique is therefore easily adaptable to perform remote measurement with optical fibres. There is very little penalty in using optical fibres, as the lasers can be launched with high efficiency, and because a fibre-based single-spatial-mode receiving filter will not lose any more light energy than the alternative of a tiny hole in a metal plate. The fibre, in fact, makes a near ideal spatial mode filter.

Specialized optoelectronics and signal processing methods for spectroscopy

This section will look at specialized spectroscopic methods such as (e.g. scanned filters, modern fixed detector arrays and use of Hadamard and Fourier transform signal processing methods. In simple dispersive spectrometer instruments, the desired selection of optical wavelength or frequency is achieved by monochromators using a prism or diffraction grating, with the necessary collimation and re-focussing optics. The spectrum is then recorded sequentially in time, as the frequency or wavelength of the light transmitted by the filter is scanned.

An alternative possibility, discussed in this section, is to use various parallel multiplexing techniques, where all wavelengths are monitored at the same time. There are two generic lines of development that have gained success in recent years.

The first of these, which has already been mentioned above, is simply to use multiple detectors, either a discrete-multi-element photodiode array with separate amplifiers, or a self-scanned CCD detector array. Both of these enable the parallel detection of all wavelengths, so more efficient use of the available light. These components were discussed above.

The second generic approach is to pass the light through a more complex multi-wavelength optical filter, which is capable of passing many wavelengths at once, but where the spectral transmission varies as it is scanned with time. Light then finally impinges on a single optical detector. The transmitted spectrum is then finally decoded by applying mathematical algorithms to the observed temporal variations in the detected signal as the complex filter is scanned. Two common variations of this so-called *transform* approach are used, first, the Hadamard and second, the Fourier method. Both methods have the advantage of parallel multiplexing, thereby achieving a co-called *Fellget* advantage [16]. Both methods have the additional advantage of high throughput [17], since a single narrow slit is no longer used, but a large area (normally a circular hole) allows far more light to enter.

In both of these transform methods, mathematical analysis is used to decode optical signals that have arisen from superposition of radiation at many different wavelengths. In the Hadamard spectrometer, an encoded mask, with a transmissive (or reflective) pattern of markings, is positioned in the focal exit plane of a normal dispersive spectrometer and then a mathematical transformation called the *Hadamard matrix* is applied. The coded mask usually has a fine-detail binary orthogonal code on it, forming a pattern of elements like 0100101001101110010 10, such that the detected transmitted (or reflected) signal varies with the position of this mask [18]. Using the Hadamard matrix to perform the desired mathematical transform, the transmitted spectrum can be reconstructed.

Fourier transform spectrometers require a filter having a sinusoidal variation of optical transmission, which can have its spectral period (wavelength or frequency between successive peaks in transmission) modulated with time in a defined manner. All 2-path optical interferometers, as their path difference is varied with time, conveniently demonstrate the desired sinusoidal transmission changes, which are a natural consequence of the interference process. The most convenient form of this to use in instruments is the Michelson interferometer configuration [19, 20].

The simplest way to appreciate how a Fourier transform spectrometer operates is to first consider what will happen with monochromatic light (e.g. single-frequency laser). Here, if the path difference of the interferometer is increased linearly with time, the intensity transmitted by the interferometer will vary in a purely sinusoidal manner with time, i.e. the output from the optical detector will be a single electronic frequency. Any more complex optical spectra can be considered to be composed of a superposition of many such pure optical frequencies and each of these single-frequency components will generate a pure electronic frequency or tone, so will give its own unique electronic signature. The temporal signal from the detector will be composed of a linear superposition of all such signals. Decoding of such an electronic signal, to recover all such single-frequency sinusoidal signal components, is a standard problem in electronic spectrum analysers and the well-known solution is that of *Fourier analysis*, hence the use of the words *Fourier transform spectrometer* to describe its optical equivalent.

In all of these transform methods, the instantaneous condition of the complex optical filter is known from the position of the interferometer (Fourier method) or the coded mask (Hadamard method), thus allowing decoding of detected signals to produce the desired spectral output.

Before finishing this section, it is instructive to discuss the relative strengths and weaknesses of each approach. In fundamental signal to noise ratio terms, it is preferable to use dispersive optics with a fully parallel array detector, as for example used to great advantage in modern CCD array spectrometers. Unfortunately, such detectors are only available in the visible and near infrared region, as detector technology for other regions is poor. Because of this, there is far greater use of Fourier transform methods in the mid- and far-IR regions, where it is only necessary to have one (often cooled) high-performance single detector. Then the Hadamard and Fourier methods gain advantage over using a single detector with a scanned narrowband filter, as more wavelengths are transmitted through the filter at any one time, and a greater optical throughput can be used in the optical system. The Fourier transform method has a further advantage, however, when high spectral resolution is desired, as it is difficult, particularly in the IR region, to make detector arrays with a small enough spacing to resolve very closely spaced wavelengths, which would otherwise need very large dispersive spectrometers to achieve high resolution.

C4.1.4 Case studies of spectroscopic methods

In the following sections, we shall now discuss several case studies, which give examples of recent research on optical sensing, and for convenience, several of these are from the work of the team of Prof. Gauglitz. Some of the descriptions will require the introduction of some further basic ideas, so the reader can best appreciate the methods used.

C4.1.4.1 Guided radiation and its use to detect changes in the real or imaginary (absorption) part of refractive index

Waveguide theory was discussed in detail in Chapter A2.1 and evanescent fields were discussed earlier in this one, but here we briefly review how waveguides can detect samples in the environment close to the waveguide, by observing their influence on the evanescent field. It is often desired to monitor changes in absorption or refractive index in a medium. Quite a few methods to interrogate changes in the real or imaginary part of the refractive index in the environment near a waveguide are known. With evanescent field methods, the field is usually accessed at some selected point along the fibre and then it is only necessary to monitor the fibre transmission.

An early method for refractive index monitoring, by Lukosz [21] (see figure C4.1.14), was to embed a grating on the surface of a slab (rectangular cross-section) waveguide, taking advantage of the modification of the Bragg grating condition when the index of the external medium changes. The original concept has since been modified by many scientists. The effective refractive index that is seen by light in the waveguide depends not only on the index of the medium close to the interface, but also on the

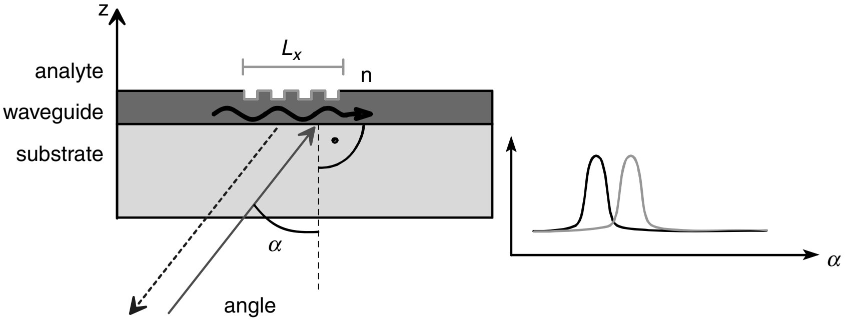


Figure C4.1.14. Grating coupler in its originally devised form [21]. Coupling is at a maximum for an angle of incidence at a Bragg angle, which depends on the refractive index of the medium next to the waveguide.

angle of incident (interrogating) radiation α, the grating constant g, the wavelength λ and the grating order m:

$$n_{\mathrm{eff}} = n_{\mathrm{env}} \sin\alpha + m\,\lambda/g \tag{C4.1.10}$$

To avoid complex alignment, a monochromatic light source (such as shown in figure C4.1.15, using a laser) can be combined with a CCD receiving array [22]. This allows measurements without any need for sophisticated adjustable mechanical alignment systems, such as goniometers.

Another, very interesting approach has been introduced by R. Kunz [23, 24], using embossed polycarbonate gratings, which can be produced in a very simple way, as shown in figure C4.1.16. Cheap disposable chips can be produced and the properties can be easily graded, as a function of distance across the device, either by modifying the groove spacing of the grating perpendicular to the waveguide structure or the depth of the waveguide below the grating. By varying either of these, in a manner that changes (tapered separation) along the out-coupling grating, variations of refractive index can be detected. By means of this simple trick, there is no need for either a goniometric approach or angle-selective measurements. A simple linear detector array will determine the position along the grating at

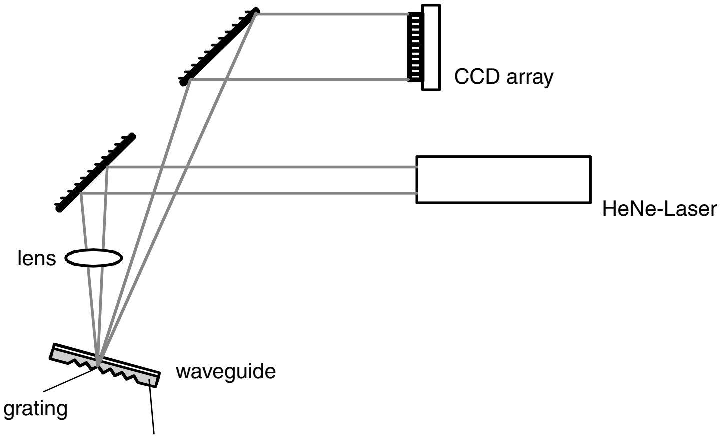


Figure C4.1.15. Grating coupler arrangement, which avoids the need for complex goniometric positioning by allowing light diffracted at different angles to be incident on different pixels on a diode array.

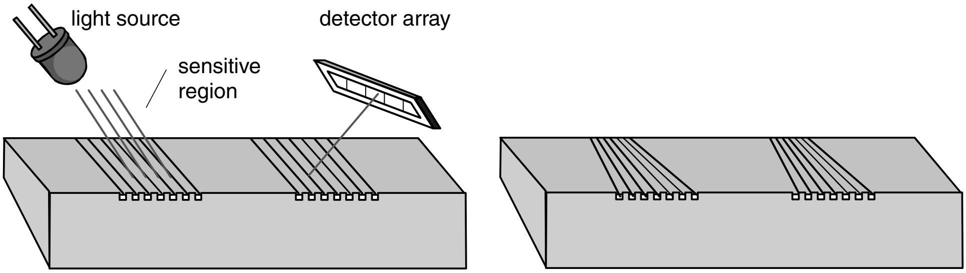


Figure C4.1.16. Grating couplers using spatial gradients of grating line spacing or of depth of grating below surface. The diagram shows the use of either variation in thickness of the waveguide or non-parallel groves.

which optimum out-coupling has occurred and this of course depends on the refractive index in the environment of the waveguide.

Figure C4.1.17 shows another approach using a prism coupler, which is a common means of coupling light into slab waveguides. This technique is also sometimes called a resonant mirror method. A layer of lower refractive index is placed as a buffer layer between the prism and the waveguide. An incident beam (polarized at 45° to the normal in the plane of incidence) will be reflected at the base of the prism in a manner dependent on the wavelength, the angle of incidence and the optical properties of the prism and/or the waveguide. The incident beam can excite TE (transversal electric) and/or TM (transversal magnetic) modes of the waveguide and the modes of the waveguide can re-couple into the prism, resulting in a phase shift relative to the directly reflected beam. Because both propagating modes (TE and TM) travel with different velocities within the waveguide and are differently influenced by the refractivity of the medium in the evanescent field region [25], the plane of polarization of the output light is in general elliptical, with a polarization state depending on the relative phase delay. The process is similar to the polarization changes that occur in a bi-refringent crystal.

Direct interferometers, i.e. ones not making use of polarization changes, are also interesting and make useful arrangements for interrogation of refractive index. In a commonly used configuration, radiation is guided via two arms of a Mach–Zehnder interferometer (see figure C4.1.18 [26]) and one of the arms is covered by a sensing layer. This is typically a polymer sensing film, but it may also be a more complex bio-molecular recognition layer. Guided radiation propagates within these two arms with different phase velocities, resulting in a phase shift that, after interferometric superposition at the

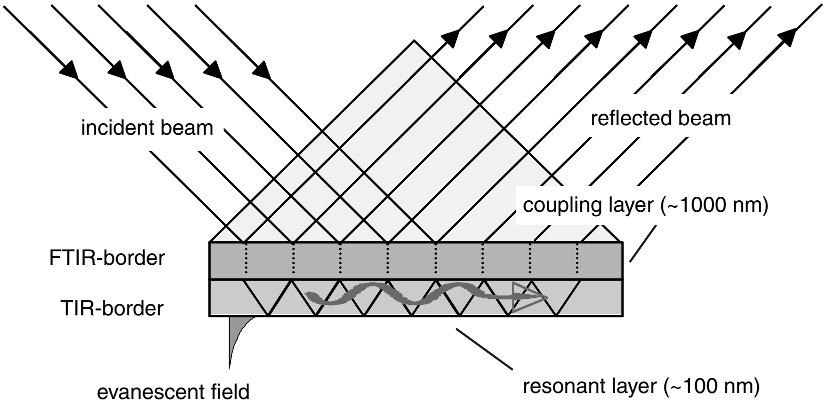


Figure C4.1.17. Principle of the prism coupler (note: sometimes called a resonant mirror).

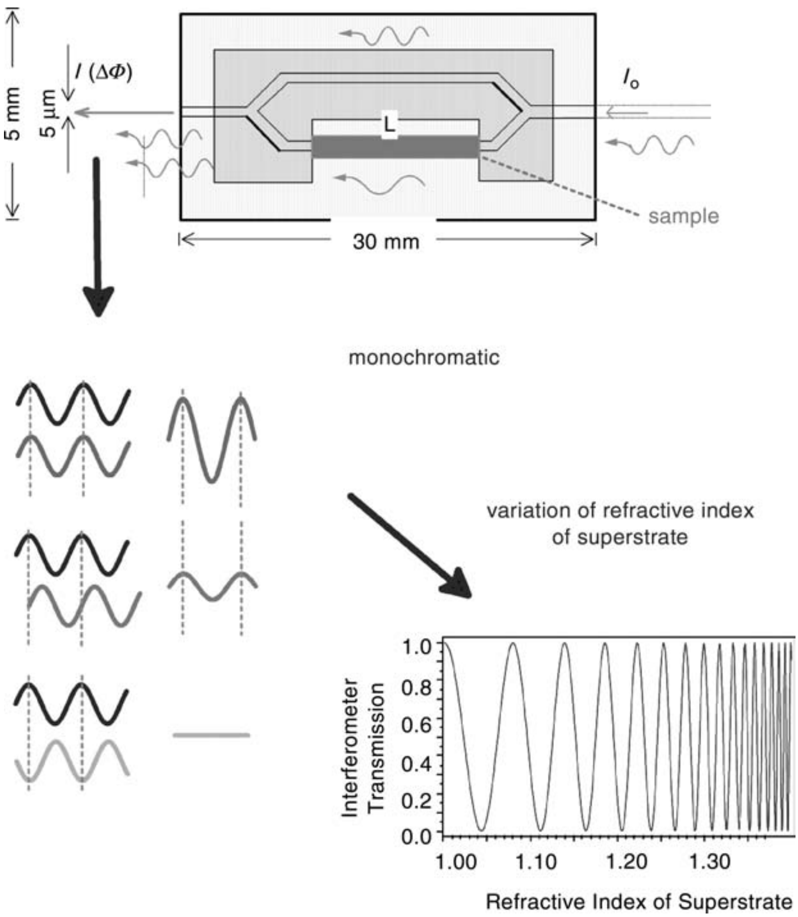

5 mm
5 µm
$I(\Delta\Phi)$
I_o
L
sample
30 mm
monochromatic
variation of refractive index of superstrate
Interferometer Transmission
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.10
1.20
1.30
Refractive Index of Superstrate

Figure C4.1.18. Depiction of the two-arm Mach–Zehnder sensing interferometer. The variations of intensity response (lower right) due to changes of refractive index within the refractive-index sensitive layer (superstrate) are shown. The superposition of partial beams is shown on the left, with different relative phases.

coupling junction, can be measured as an intensity change, which is now dependent on the refractive index of the medium in contact with the recognition layer. The phase of the beams, and hence the signal intensity, depends, of course, on the optical wavelength, the interaction length, and the waveguide properties. The phase shift-dependent intensity is given by equation C4.1.11 below, where L is the interaction length, Δn_{eff} is the change in the effective refractive index within the waveguide arm (such changes caused by changes in the evanescent field region), and k is the wave vector.

$$I(\Delta\Phi)/I_0 = 1/2[1 + \cos(Lk\Delta n_{\mathrm{eff}})] \tag{C4.1.11}$$

In figure C4.1.18, the interference pattern is shown as a superposition of the two propagating waves in the two different waveguide arms.

Another possible configuration is the well-known Young interferometer [27] arrangement. Instead of recombining the two waveguides in the planar layer, the two beams are now focussed in one plane only, using a cylindrical lens, and directed onto a CCD array, where they form an interference pattern, as shown in figure C4.1.19.

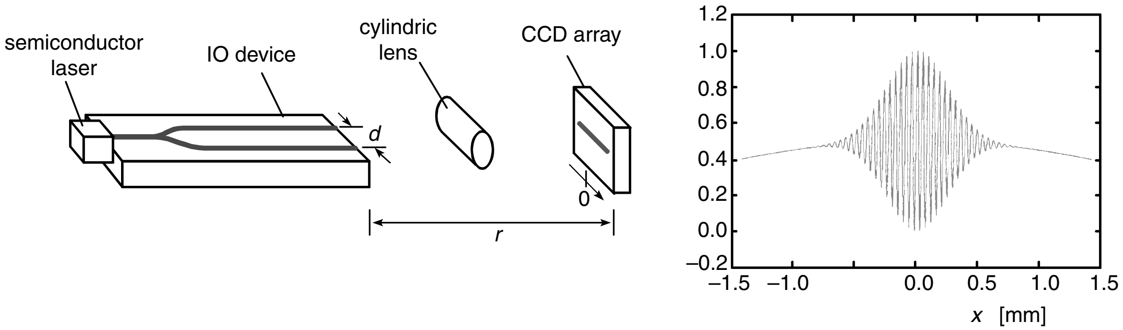


Figure C4.1.19. Young's interferometer: this shows a waveguide version of the interferometer and spectral detection of the superimposed beams and interference pattern in the far field. (A cylindrical lens improves the signal intensity by reducing spread due to diffraction in the vertical direction.)

Surface plasmon resonance is a means of enhancing the evanescent field intensity using thin metal films. The surface plasmons (electrons) in a metal film on its surface can be excited by a guided wave in the waveguide. The excitation has a dependence on the refractive index of the medium on top of the metal film, particularly at the resonance condition, where the intensity of the radiation propagating in the waveguide is substantially reduced by the stronger resonance coupling [28]. It can be used as a detection principle, either directly, using a waveguide, or indirectly, using a waveguide structure in combination with a prism. One arrangement, where a slab waveguide is coated with a buffer layer and a metal film, is shown in figure C4.1.20.

Originally, surface plasmon resonance was introduced to the scientific community by Kretschmann [29] and Raether [30]. Two approaches are commonly used. The first uses monochromatic light with a fixed angle of incidence, and then the resonance condition, which determines where light falls on a position-sensitive linear photodiode array (see figure C4.1.21*a*), is monitored [31]. At present, this is one of the most commonly used optical methods for interrogating bio-molecular systems. It has been commercialized by Biacore for examination of affinity reactions (www.biacore.com).

Another approach is shown in figure C4.1.21*b*, where the optical system is arranged so that it receives only a narrow angular range of light. With this system, the need to meet the resonance condition causes a narrow range of wavelengths from a broadband white light source to be selectively attenuated [32]. Using a diode-array spectrometer, it is then possible to record the wavelength of the 'dip' in the spectrum of the internally reflected signal.

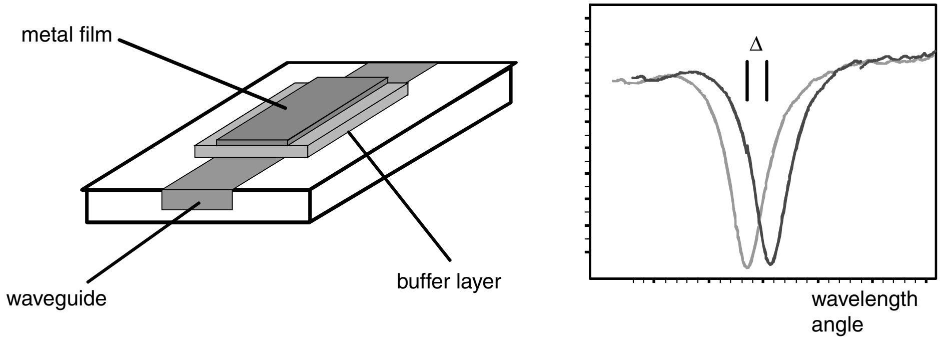


Figure C4.1.20. Waveguide-based surface plasmon resonance.

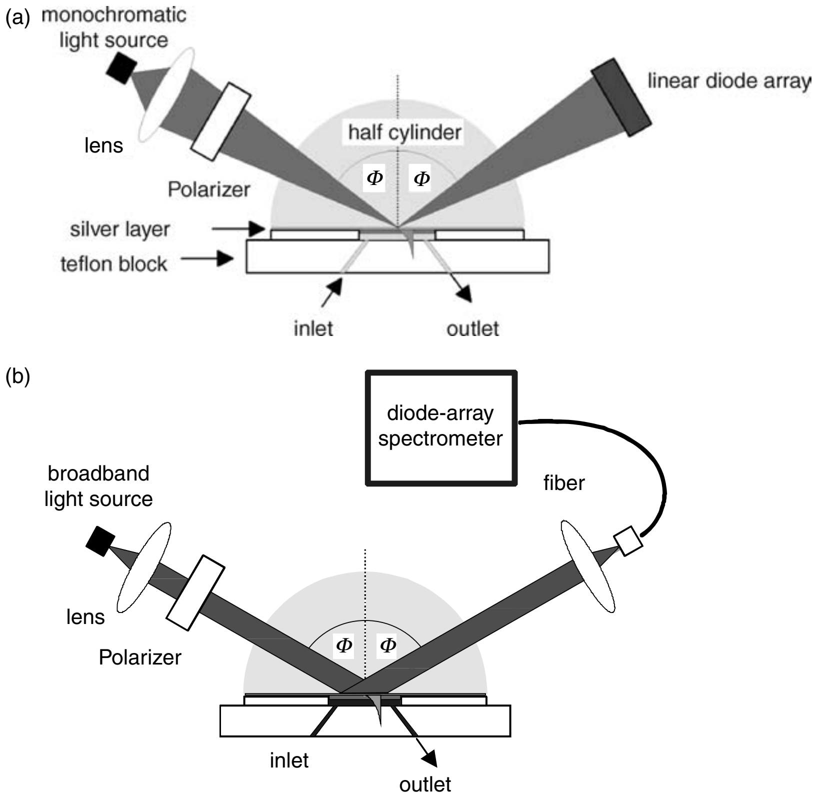

(a)
monochromatic light source
lens
Polarizer
half cylinder
Φ
Φ
linear diode array
silver layer
teflon block
inlet
outlet
(b)
diode-array spectrometer
fiber
broadband light source
lens
Polarizer
Φ
Φ
inlet
outlet

Figure C4.1.21. Surface plasmon resonance. (A) Monitoring of angular dependence, with a narrowband light source. (B) Monitoring of wavelength dependence, at fixed angle, with white light input and spectral readout, using a diode-array spectrometer.

In addition to the above examples of interrogation principles, many other approaches have been published [33]. Recent publications review more details of new developments and applications of these methods [34–38].

C4.1.4.2 Multiple reflections, interference

Apart from refractometry, reflectometry is frequently used in sensing. For this reason, a brief introduction to white light interferometry will be given below giving an example of an approach similar to the Fabry–Perot interferometer. For many decades, *ellipsometry* has been known to examine the properties of thin film layers. As mentioned above, reflection and refraction takes place at each interface between media of different refractive index. The partially reflected beams from the interfaces at each side of a layer will superimpose and exhibit interference. The intensity of reflection depends on the wavelength, the angle of incidence of radiation, the refractive index of the layer and the physical thickness of this layer. In *ellipsometry* [34, 35] polarized light is used and thus the refractive index and the physical thickness of the layer can be determined separately. In normal reflectometry, no polarization state is selected, resulting in a very simple, robust and easy-to-use method for monitoring effects in these layers. As can be seen in figure C4.1.22, the two partial beams can show constructive or destructive interference or any state between, depending on the above mentioned parameters. In this figure, the path

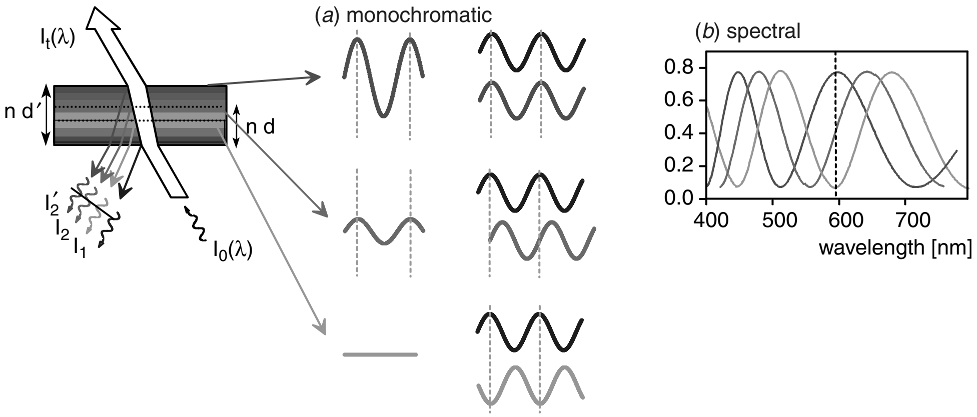


Figure C4.1.22. Reflectometric interference spectroscopy (RIfS) [36, 37], where the superposition of partial beams reflected at the interfaces of a thin layer is measured. Any superposition causes constructive and destructive interference, resulting in a wavelength dependence. When white-light incident radiation is used, this results in the interference spectrum given on the right side of the diagram. If just one reflected wavelength (monochromatic) from the layer is measured, the intensity of the reflected radiation varies in a periodic manner as the optical thickness of the layer (i.e. change in thickness or refractive index) is changed. This is shown by the dotted vertical line in the diagram.

of light is shown at an angle in order to be able to visualize the different partial beams but the light beam will normally be at an angle close to normal to the surface.

For the simple case of a non-absorbing layer, having multiple reflections at only two interfaces, equation C4.1.12 describes the intensity modulation with wavelength. Changes in the properties of the layer, particularly of a polymer layer, can be induced by adsorption of analytes such as gases and liquids. These can change the effective optical thickness of the layer, by varying either the refractive index or the physical thickness of this layer (swelling). Adsorption into this layer is most likely in the case of polymer films. In order to monitor biomolecular interactions, the effects of various affinity reactions can also be observed.

$$I_\lambda = I_1 + I_2 + 2\sqrt{I_1 \cdot I_2}\,\cos\left(\frac{4\pi n d}{\lambda}\right) \qquad \text{(C4.1.12)}$$

C4.1.4.3 Experimental apparatus for the RIfS method

Figure C4.1.23 shows the two different approaches for measuring gases or liquids. These have the combined advantages of small cell volume and the possibility of remotely monitoring interference effects in the cell via fibre optic leads.

Sample preparation is an important aspect of spectroscopy and examples of arrangements for this for both gas and liquid are shown in figure C4.1.24. In the case of gases, a flow controller is usually used, which allows multi-analyte detection. In the case of liquid samples, a flow injection analysis system containing pumps is usually used (preferably syringe pumps) along with selection valves.

Monitoring the reflection from Fabry–Perot polymer layers involves cheap and simple layers that can be made from many different thicknesses, to enable sensing arrays to be constructed. By monitoring the response of many different layers/materials to the same chemical influence, a different response

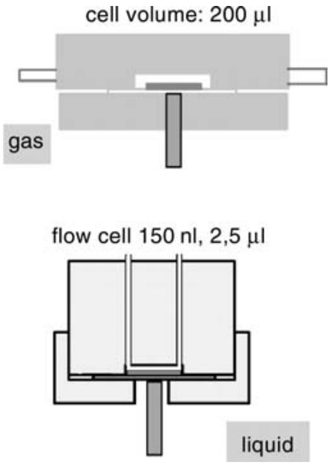

cell volume: 200 µl
gas
flow cell 150 nl, 2,5 µl
liquid

Figure C4.1.23. Cell compartment for measuring gases and liquids, using RIfS.

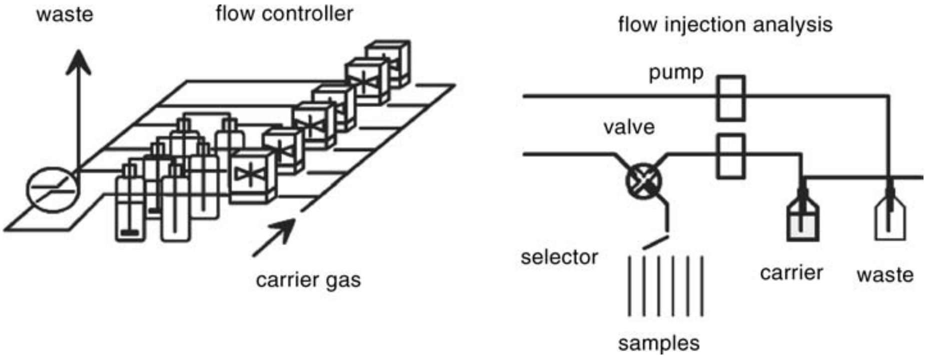

waste
flow controller
carrier gas
flow injection analysis
pump
valve
selector
samples
carrier
waste

Figure C4.1.24. Sample preparation is achieved either by flow controllers (for gas) or by a flow injection analysis setup (liquids).

pattern to each individual constituent can be measured and, when combined with a pattern recognition system, the arrangement can make a simple, but effective, form of opto-electronic nose.

C4.1.5 Conclusions

We have now completed our simple introduction to the subject of spectroscopy, in order to give an overview to non-specialist readers. This area is sufficiently complex to fill large specialist textbooks, so it is hoped the reader will forgive us where there are inevitable omissions. We have tried to give a short introduction to the basic theory, a brief practical overview of components and instruments and finally introduce a few examples of recent research areas, mainly using fibre optics. Clearly the interested reader can gain more insight into this fascinating subject from what are usually more-voluminous specialist textbooks, many of which have been mentioned in the references and further reading.

References

[1] Naumer H, Heller W and Gauglitz G (eds) 2003 *Untersuchungsmethoden in der Chemie* chapter 8 (Weinheim: Wiley-VCH)

[2] Gauglitz G 1994 Ultraviolet and visible spectroscopy *Ullmann's Encyclopedia of Industrial Chemistry* Vol. B5 (Weinheim: Wiley-VCH)

[3] Ingle J D Jr and Crouch S R 1988 *Analytical Spectroscopy* (Englewood Cliffs, NJ: Prentice-Hall)

[4] Svehla G (ed) 1986 Analytical visible and ultraviolet spectrometry *Comprehensive Analytical Chemistry* Vol. XIX (Amsterdam: Elsevier)

[5] Skoog D A and Leary J J 1992 *Principles of Instrumental Analysis* (Fort Worth: Saunders College Publishing)

[6] Born M and Wolf E 1980 *Principles of Optics* (New York: Pergamon)

[7] Hecht E and Zajac A 1974 *Optics* (Reading: Addison-Wesley)

[8] van der Hulst H C 1957 *Light Scattering by Small Particles* (New York: Wiley)

[9] Stacey K A 1956 *Light-Scattering in Physical Chemistry* (London: Butterworths)

[10] Harrick N J 1979 *Internal Reflection Spectroscopy* (New York: Harrick Scientific Corporation)

[11] Naumer H, Heller W and Gauglitz G (eds) 2003 *Untersuchungsmethoden in der Chemie* chapter 15 (Weinheim: Wiley-VCH)

[12] Klessinger M (ed) 1995 *Excited States and Photochemistry of Organic Molecules* (New York: John Wiley & Sons Inc.)

[13] Wolfbeis O S (ed) 1991 *Fibre Optic Chemical Sensors and Biosensors* (Boca Raton: CRC Press)

[14] Narayanaswamy R and Wolfbeis O S 2004 *Optical Sensors* (Berlin, Heidelberg: Springer Verlag)

[15] Colthup N B, Daly L H and Wiberley S E 1990 *Introduction to Infrared and Raman Spectroscopy* (New York: Academic Press)

[16] Brown R G W and Smart A E 1997 *Appl. Optics* **36** 7480–7492

[17] Fellgett P B 1958 *J. Phys. Radium* **19** 187–191

[18] Jacquinot E 1954 *J. Opt. Soc. Am.* **44** 761–765

[19] Griffiths P R 1975 *Chemical Infrared Fourier Transform Spectroscopy* (New York: Wiley)

[20] Griffiths P R and De Haseth J A 1986 *Fourier Transform Infrared Spectrometry* (New York: Wiley)

[21] Nellen P h M and Lukozs W 1993 *Biosens. Bioelectron.* **8** 129–147

[22] Brandenburg A and Gombert A 1993 *Sens. Actuators B* **17** 35–40

[23] Kunz R E, Edlinger J, Curtis B J, Gale M T, Kempen L U, Rudigier H and Schuetz H 1994 *Proc. SPIE-Int. Soc. Opt. Eng.* **2068** 313–325

[24] Kunz R E 1991 *Proc. SPIE-Int. Soc. Opt. Eng.* **1587** 98

[25] Cush R, Cronin J M, Stewart W J, Maule C H, Molloy J and Goddard N J 1993 *Biosens. Bioelectron.* **8** 347–354

[26] Ingenhoff J, Drapp B and Gauglitz G 1993 *Fresenius J. Anal. Chem.* **346** 580–583

[27] Brandenburg A and Henninger R 1994 *Appl. Opt.* **33** 5941–5947

[28] Piraud C, Mwarania E, Wylangowski G, Wilkinson J, O'Dwyer K and Schiffrin D J 1992 *Anal. Chem.* **64** 651–655

[29] Kretschmann E 1971 *Z. Phys.* **241** 313–324

[30] Raether H 1977 *Phys. Thin Films* **9** 145–261

[31] Liedberg B, Nylander C and Lundström I 1983 *Sens. Actuators B* **4** 299–304

[32] Brecht A, Gauglitz G and Striebel C 1994 *Biosens. Bioelectron.* **9** 139–146

[33] Gauglitz G 1996 Opto-chemical and Opto-immuno sensors *Sensors Update* Vol I (Weinheim: VCH Verlagsgesellschaft mbH)

[34] Liebermann T and Knoll W 2000 *Colloids and Surfaces A* **171** 115

[35] Anton van der Merwe P 2001 Surface plasmon resonance *Protein–Ligand Interactions: Hydrodynamics and Calorimetry* (Oxford: Oxford University Press)

[36] Kinning T and Edwards P 2002 *Optical Biosensors* (eds) F S Ligler, T Rowe and A Chris (Amsterdam: Elsevier)

[37] Kuhlmeier D, Rodda E, Kolarik L O, Furlong D N and Bilitewski U 2003 *Biosens. Bioelectron.* **18** 925

[38] Klotz A, Brecht A, Barzen C, Gauglitz G, Harris R D, Quigley Q R and Wilkinson J S 1998 *Sens Actuators* **B51** 181

Further reading

Dyer S A *et al* 1992 Hadamard methods in signal recovery *Computer-enhanced Analytical Spectroscopy* Vol. III (ed) P C Jurs (New York: Plenum Press)

Azzam R M A and Bashara N M 1989 *Ellipsometry and Polarized Light* (Amsterdam: North Holland)

Arwin H 2003 Ellipsometry in life sciences *Handbook of Ellipsometry* (eds) G E Jellison and H G Tompkins (Park Ridge, NJ: Noyes Publications)

Gauglitz G and Nahm W 1991 *Fresenius.Z Anal Chem.* **341** 279–283

Brecht A, Gauglitz G, Kraus G and Nahm W 1993 *Sensors and Actuators B***11** 21–27

Brecht A, Gauglitz G and Göpel W 1998 Sensor applications *Sensors Update* Vol. III (eds) H Baltes, W Göpel and J Hesse (Weinheim: Wiley-VCH)

Gauglitz G Optical sensors: Principles and selected applications, *Anal. Bioanal. Chem.* **381** 141–155

Smith B C 1995 *Fundamentals of Fourier Transform Infrared Spectroscopy* (Boca Raton: CRC Press)

Mark H 1991 *Principles and Practice of Spectroscopic Calibration* (New York: John Wiley)

Robinson J W (ed) 1974 *CRC Handbook of Spectroscopy,* Vols I–III (Ohio: CRC Press)

C4.2
Intelligent surveillance

Brian Smith

C4.2.1 Introduction

Remote surveillance in the form of closed circuit television (CCTV) has been available almost since the invention of television. It has the great advantage in providing a remote 'eye' to capture the scene, which may often be in an environment hostile to a viewer. However, a viewer is still required to analyse the data, and one may be looking for sparse events, such as an intruder breaking into a building. Images however can now easily be captured in a digital format which in turn can be subject to algorithmic analysis and search techniques in a computer. This leads to automation of the analysis and viewing burden, and is often described as 'intelligent surveillance'. The real benefit, however, often derives as much from the automation of the decision-making task as the analysis applied to the images.

Applications of image processing for intelligent surveillance include ocular iris or fingerprint matching, facial recognition and intruder detection. This chapter is in the form of a case study of automatic license plate reading (ALPR), which illustrates virtually every aspect of the general problem, with a general discussion of applications of digital CCTV following.

C4.2.2 Fundamentals

Each intelligent surveillance system typically breaks down into a number of sub-systems.

These typically include:

- *Image generation*. The greatest returns on investment come from engineering the scene, the illumination and/or the sensor technology so that the object(s) of interest are readily distinguishable in the image, and the irrelevant detail is suppressed as far as possible. It may well be possible to engineer an overall control loop whereby the scene is adjusted by analysing a metric on the detected object, e.g. the illumination is changed dynamically to improve the contrast between the object of interest and its background.
- *Image digitization*. High-speed digital circuits have made this a readily obtainable off the shelf solution.
- *Feature detection*. Typically there are time constraints to the amount of computing that can be applied to a recognition problem. The effect of the computing can greatly be enhanced if it is concentrated on the features of interest rather than the scene in general. The art of pattern recognition is generally defined through the choice of the best feature set that can be extracted to adequately and accurately describe the object of interest in the time available. Some features like the average contrast or average brightness of the scene are trivial to compute.

- *Feature-set classification*. This is the classification of the feature vector into objects of interest, e.g. the objective may be to recognize printed characters. The pixels in the image are reduced to features. The features should be well chosen so that when the vector of features associated with one object clearly separates it from the vector associated with another. For example, there might be a feature which is the presence of the stroke in a 'Q', which seeks to separate it from an 'O'. Classification in this case would be making a decision on the presence or absence of the stroke feature.
- *Decision outcome*. Real systems will have an action which is dependent on the outcome of the classification, e.g. to raise an alarm on the detection of smoke, or to raise a barrier upon reading a permitted vehicle license plate.

C4.2.3 Example systems

C4.2.3.1 Automatic license plate reading

The task of license plate reading is generally made easier through the adoption of state-wide and national, even international standards for vehicle license plates. The format of the license plate is thus generally well-defined. However, the environmental conditions under which one is seeking to do so are very harsh. Salt and road-dirt can mask the letters; one has to cope with vehicles travelling at a wide variety of speeds, and possibly close together; the weather and lighting conditions can vary enormously. Typically, 24/365 operation is required, and the system should not distract the driver in any way.

Image generation

Freezing vehicle motion

The standard exposure time for CCTV cameras is 20 ms in Europe (16.7 ms, USA). During this time, a vehicle travelling at 100 kph (60 mph) will move approximately 0.5 m (18 in) and the license plate will be unreadable. To reduce the blurring caused by the movement of the vehicle, a much shorter exposure time must be used, typically 1 ms or less. CCD-based CCTV cameras achieve this by means of 'electronic shuttering'. The picture information falling on the sensor results in an electrical charge that is accumulated for the first 19 ms (15.7 ms, USA) of the field and is then discarded by dumping it into the substrate of the sensor. Accumulation of picture information restarts at this point but can now only be for the final 1 ms of the field.

Illuminating the license plate

A 1 ms exposure time reduces the sensitivity of the camera to about 6% of its unshuttered sensitivity. Interfering light sources that operate continuously, such as vehicle head lamps, are therefore much reduced because the camera only responds to light from these sources for a short time. Unfortunately, the reflected light from continuously operating IR illuminators, such as filtered incandescent lamps, is also reduced by the same proportion and a more powerful IR illuminator will be required to ensure adequate picture quality. It is true to say that with this type of illuminator 94% of the light is wasted.

A novel solution to this problem has been devised [1]. Many types of infrared-light-emitting diode (IR-LED) can be pulsed to high current levels for short periods of time. With suitable drive circuitry, the IR energy that would be emitted by the LEDs during the normal 20 or 16.7 ms camera exposure time can be compressed into the 1 ms exposure time of the shuttered camera. Consequently, the camera will still receive substantially the same IR energy from the pulsed IR-LED illuminator as it would have received if unshuttered, but with the interfering effects of the headlights reduced to about 1/20. The IR-LED approach has also enabled the camera-illuminator to be classified as eye-safe with a nominal ocular hazard distance (NOHD) of 0 m, without magnification [2–4]. This is in sharp contrast to other forms of illumination such as filtered halogen, where the NOHD is much greater.

Overcoming headlights

The effect of headlights can be reduced still further. IR-LEDs emit most of their power over a relatively narrow band of wavelengths, typically 60 nm wide, whilst incandescent head lamps emit over a band of wavelengths from UV through visible into the middle IR part of the spectrum (a waveband from 350 to 2700 nm, filtered by the head lamp glass). Figure C4.2.1 compares the measured spectra of a quartz-halogen lamp and a 950 nm IR-LED. By incorporating a suitable optical bandpass filter between the lens and the CCD the majority of the light from head lamps can be filtered out. This gives a further factor of ten reduction of light from the head lamp relative to the reflected light from the IR-LED illuminator.

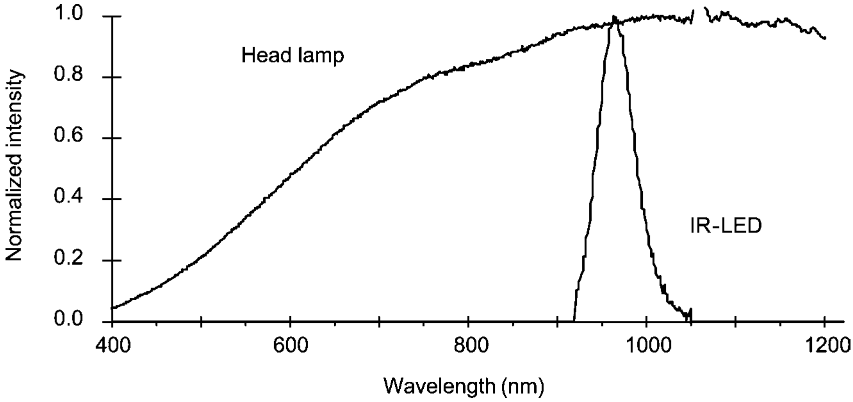


Figure C4.2.1. Head lamp and IR-LED spectra. (Courtesy of PIPS Technology Ltd.)

By combining pulsed IR illumination with optical filtering, the effect of head lamps can be reduced by about 200 times relative to the light reflected from the license plate. Furthermore, if the illuminator and camera are co-aligned, the high reflection capability of retro-reflective license plates, which are mandatory in many countries, may be used. The IR power required, and hence the overall illuminator power, is minimized. The illuminator can be fitted around the camera lens within a single housing, and the P362 traffic camera shown below consumes a maximum electrical power of 15 W. The low power enables mobile systems to be deployed, e.g. mounted on a police vehicle. Furthermore, the costs of electricity for a continuously operated high power illuminator are considerable and lamp changes are required every few months. The P362 camera also includes a second wide angle ‘overview’ colour camera within the same housing to capture a contextual image of the vehicle (figure C4.2.2).

Daytime operation

The high current pulsing of the IR-LED illuminator and spectral filtering of the camera are also effective against sunlight. The sun can be approximated to a black body radiator at a temperature of about 5800 K but the spectrum is modified by absorption by atmospheric gases, predominantly water vapour. Figure C4.2.3 shows the spectrum of sunlight measured in central England in September 1996.

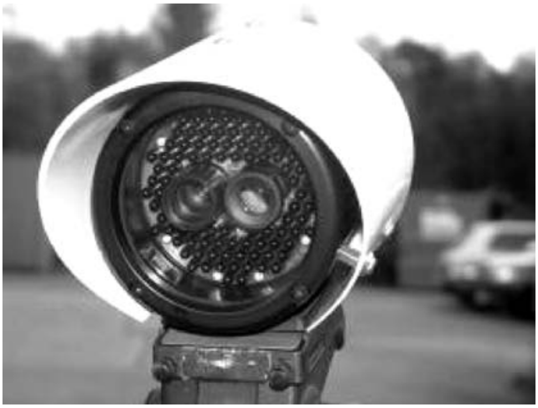


Figure C4.2.2. P362 camera with integral illuminator and overview camera. (Courtesy of PIPS Technology Ltd.)

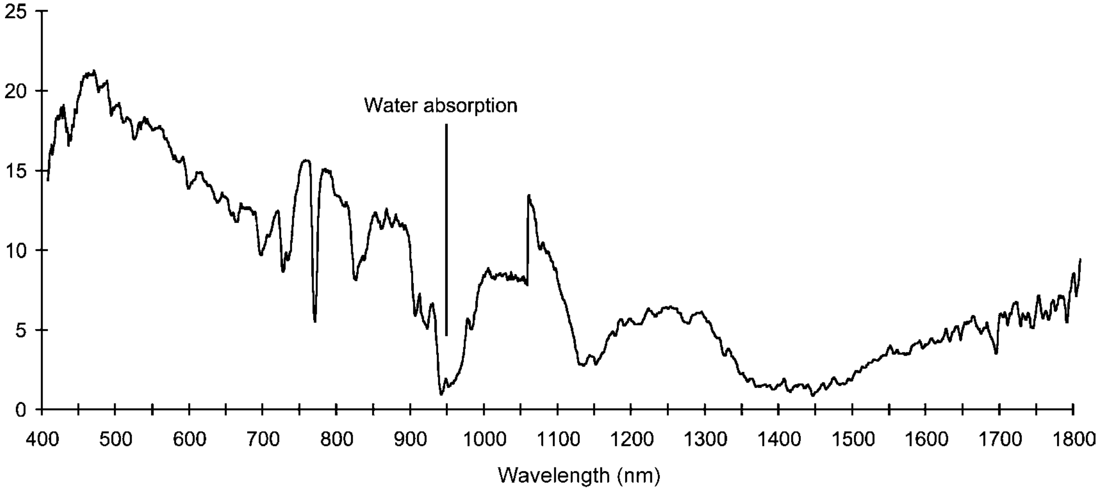


Figure C4.2.3. Spectrum of sunlight showing H_2O absorption in vicinity of 950 nm. (Courtesy of PIPS Technology Ltd.)

As can be seen, there is an atmospheric 'hole' with a centre wavelength of 950 nm and with a width similar to the emission spectrum of a 950 nm IR-LED. Thus atmospheric water vapour adds significantly to the suppression of sunlight during daytime operation. The depth of the 'hole' at 950 nm is dependent upon the amount of precipitable water vapour in the atmosphere, but is a nonlinear function in that the initial few millimetres of precipitable water deplete as much energy as the last few centimetres [5].

Overcoming variations in license plate reflectivity

A film of dirt on the surface of the plate degrades the reflection factor of retro-reflection license plates. The variability between clean and dirty plates can be countered by changing the energy of the illumination pulse on a field-by-field basis. For example, a repeating sequence of three illumination pulses, high, medium and low energy, can be used so that a clean plate that might overload the camera

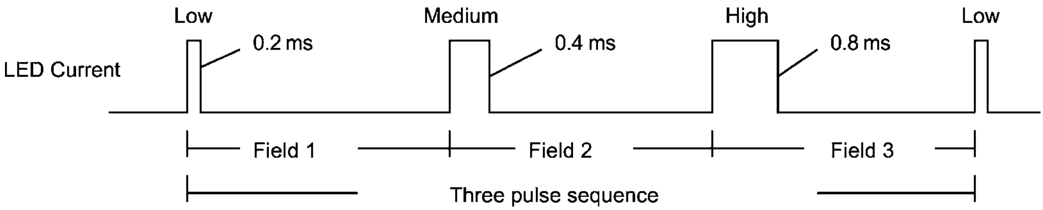


Figure C4.2.4. Three energy flash sequence. (Courtesy of PIPS Technology Ltd.)

on medium or high energy can be read during the low energy field, whilst a dirty plate which gives an unusable image on the low energy field will give a much better image on the medium or high energy fields. The energy can be varied either by changing the current through the LEDs or by changing the duration of the pulse as shown in figure C4.2.4 — the shorter the pulse the less the energy.

Depending on the deployment constraints of the camera (e.g. from a motorway bridge), high speed traffic may only be present within the field of view of the camera for a few fields, so it is vital that different illuminations are tried on an a priori basis rather than seeking to respond to the image content on a field by field basis. Closed loop control may be used on the settings as a whole as a 'slow loop' enabling the settings to respond to weather and lighting changes.

Digitization and feature detection

A modern high resolution video CCD has 720 pixels per line, with 288 (240, USA) video lines in the vertical direction per video field. When digitized at full resolution there will therefore be 202 kB of data per video field (assuming 1 byte per pixel). At 50 fields-per-second (60, USA), the volume of data that has to be processed rapidly is over $10\,\mathrm{MB\,s^{-1}}$. This is beyond the capabilities of all but the very fastest processors.

The problem can, however, be greatly simplified if the license plate is detected within the image by special purpose hardware, and then only the plate image or 'patch' is subject to optical character recognition (OCR) rather than the whole video field. A hardware 'plate-finder' can be used to form a thresholded measure of the spectral features of the video in real time in two dimensions [1]. A license plate can therefore be detected 'on-the-fly' as the video signal is generated. This frees up the general purpose processor to only have to deal with plate patches. License plates can therefore be detected without any trigger to say when the vehicle is present. An alternative is to have a secondary vehicle detection system, such as a high speed inductive loop buried in the road surface, detect the vehicle and act as a vehicle presence trigger to the system. Thus, the plate detection problem is removed through the use of dedicated hardware.

Classification

The patch is then segmented into characters and conventional OCR techniques, either using template matching upon a set of features extracted from the image or by means of a 'neural network' with a learning capability. Whilst a template matching approach can require more modification for a new font-set than a neural network, the absence of the need for a large plate training database is a positive advantage. Syntax checking can then be applied to the character string to resolve ambiguities, such as '0' (zero) versus 'O' (oh).

The characters on the license plate must be of a sufficient size in the image to be read reliably. With the size of characters used on UK plates (80 mm high), the scene viewed by the camera at the vehicle distance will usually be 2–2.5 m width and so a license plate capture camera will only view a single lane.

The general processing can be carried out on a PC hardware platform. However, for roadside deployment a single board running an embedded operating system, without a hard disk or fans, is much more preferable for robust operation; e.g. the P367 video recognizer shown below can handle up to four high speed streams of traffic [1] (figure C4.2.5).

Figure C4.2.5. P367 video recognizer for roadside deployment. (Courtesy of PIPS Technology Ltd.)

Decision outcomes

Automatic reading of a vehicle license plate has many applications. These include journey time measurement systems (JTMS — automatic congestion detection), policing (detecting vehicles on a ‘black-list’), toll and bus-lane enforcement (allowing permitted vehicles on a ‘white-list’, and capturing violators), speed enforcement (average speed over known distance) and access control (detecting fraud such as ticket swaps).

The largest system in the world for monitoring traffic congestion by measuring journey time is deployed by Trafficmaster Ltd in the UK [6]. Over 3500 ALPR cameras and associated video-recognizers are installed at regular intervals along major roads and highways. License plates are read at each site, shortened to protect vehicle identity and transmitted to a central location in an encrypted form as shown in figure C4.2.6.

As has been seen, great care is taken in the design of the system to extract the license plate from the surrounding scene as soon as possible in the processing chain. This enables a very cost-effective solution to be produced, and results in a system which is robust to real-world conditions and can be deployed at the roadside to minimize communications costs.

C4.2.3.2 Digital CCTV Surveillance

The advent of relatively inexpensive digital video technology has revolutionized CCTV surveillance. Critical to the design of such systems is the manipulation of large amounts of digital data. A good quality uncompressed colour image generates about 400 kB of digital data every 20 ms (16 ms if 60 Hz operation). Thus one is faced with a digital stream of over 20 MB of data per second from just a single

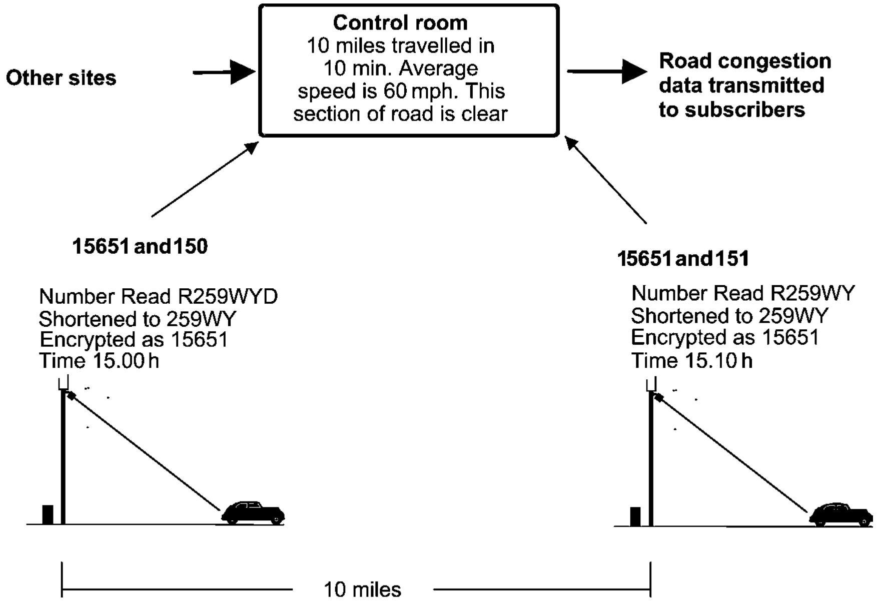

Control room
10 miles travelled in 10 min. Average speed is 60 mph. This section of road is clear
Other sites
Road congestion data transmitted to subscribers
15651 and 150
Number Read R259WYD
Shortened to 259WY
Encrypted as 15651
Time 15.00 h
15651 and 151
Number Read R259WY
Shortened to 259WY
Encrypted as 15651
Time 15.10 h
10 miles

Figure C4.2.6. Example showing two journey time measurement sites.

video source! This is within the capacity of a modern computer disk, provided there is sufficient computing bandwidth to route the stream to disk. However, communication costs become substantial at these data rates even over short distances.

There is however typically a great deal of redundant data within an image which may be compressed to reduce the data transmission bandwidth. Spatially, there will often be large areas of similar texture or colour. Since the eye is not as sensitive to high spatial frequencies as it is to low frequencies, the image may be transformed into the spatial frequency domain, and a lower number of bits used to represent it by reducing the number of bits used to quantize the higher frequencies to which the eye is not as sensitive. JPEG compression (using linearly spaced discrete cosine transforms) and wavelet compression (using octave-spaced windowed 'wavelets') is readily available on modern computers [7]. Both of these approaches allow the data volume to be traded off against image quality. Typically, compression artefacts are difficult to spot at compressions $< 3:1$, though the image quality for surveillance purposes remains practically useful at much higher compression ratios, as high as 20:1. The so-called motion JPEG (M-JPEG) is simply a stream of JPEG encoded video fields, with each video field being treated as a still image.

There is also a great deal of redundancy between successive images. Typically, there will not be very much movement within the scene from one image to the next given that there is only 20 ms between successive fields. Simple compression schemes might only transmit the differences between successive images, but they require a reference video field to be generated periodically against which the differences are computed. Video image quality may be sustained with high compression in scenes containing movement if areas of similar content are identified on a field to field basis with the amount of motion between the identical areas being transmitted rather than the areas themselves. These concepts have been embodied in the MPEG set of standards and MPEG-2 is widely available on computers.

For digital surveillance, however, a simple field-by-field encoding such as used in M-JPEG or wavelet offers many advantages. These include the encoding technology being much simpler, and the ease with which one can move around the encoded images in a real-time sense. A single JPEG or wavelet encoded video field includes all the data to re-construct it. MPEG-2 compression required the nearest reference field to be found, and all changes applied between that field and the point in question to be computed. M-JPEG fields can simply be interleaved from multiple sources into a single digital stream and easily re-constructed into separate video outputs; this is much more difficult with MPEG-2.

Compression at source enables computing communications and infrastructure to be used for intelligent surveillance. The benefits of this include:

- Direct recording to computer disk, with random access replay.
- Transmission over the Internet.
- Automatic back-up of data with original quality preserved indefinitely.
- Remote access to data from a standard PC computer.
- Encryption at source for high-security applications.
- Intelligent alarms, automatically routed via internet or mobile phone.

The area of ‘intelligent alarms’ perhaps offers the greatest scope for innovation. Automatic surveillance software will permit areas of the scene to be interactively ‘zoned’ on the displayed image and any rapid movement within the prescribed area detected. This is straightforward under controlled lighting conditions, where the natural variability within the area of interest is limited. Out-of-doors this is more difficult with lighting changes or pan/tilt camera movements being relatively rapid, and simple field-by-field differencing is likely to throw up many false alarms. However, the nature of such changes is predictable — camera movement will by definition be transitory and cause all pixels to change; changes in lighting will not change the local spatial frequency spectrum. Signal processing techniques can now economically provide solutions for many of these situations, and products are available to support ‘record-on-event’ working [8, 9]. Distinguishing between types of moving object, say between a dog and a person, may generally involve a training exercise, so that thresholds within the system are set to trigger on the person only.

The early detection of smoke or fire is now possible by video analysis [8]. The many well-known models of early combustion, including the wispy nature of smoke in the very early stages of a fire, have been characterized as images. Software can now be deployed to analyse video in real time and detect such features as wispy smoke from an electrical fault or the flicker characteristic of flames when there is little smoke present. It is also possible to distinguish between the steam naturally present in a manufacturing process, and abnormal smoke from an electrical fault [10].

C4.2.4 Conclusion

Digital video processing is making the intelligent processing of video images at source feasible for many real-world situations. Previously this required central processing with the resultant high band-width transmission of the images. Distributing the processing to the point of capture enables a few bytes of information describing the event to be transmitted where previously tens of megabytes of data were required to be transmitted per second to achieve a result.

References

[1] www.pipstechnology.com
[2] British Standard EN 60825-1:1994 Amendment 2 and IEC 825-1:1993 *Safety of Laser Products*

[3] American Conference of Government Industrial Hygienists 1996 *Threshold Limit Values and Biological Exposure Indices*
[4] Sliney D and Wolbarsht M 1980 *Safety with Lasers and Other Optical Sources* (New York: Plenum)
[5] Iqbal M 1983 *An Introduction to Solar Radiation* (New York: Academic)
[6] www.trafficmaster.co.uk
[7] Watkinson J 2000 *The Art of Digital Video* (Woburn, MA: Focal)
[8] www.issecurity.com
[9] www.pi-vision.com
[10] www.intelsec.com

C4.3
Optical actuation and control

George K Knopf

C4.3.1 Introduction

Actuators are devices that perform mechanical work in response to a command or control signal. The device can be separated into two parts: the actuator shell and the method of actuation. The shell is the basic structure of the actuator and, often, contains deformable or moving parts. Pneumatic cylinders and mechanical linkages are large-scale actuator shells. Examples of deformable microactuator shells include cantilever beams, microbridges, diaphragms, and torsional mirrors [1]. The main function of any shell design is to provide a mechanism for the desired actuation method to produce useful work. The actuation method is the means by which a control signal is converted to a force that is applied to the actuator shell and creates physical movement. The output of the overall system is the desired response given as a displacement, force, or pressure value. The different methods of actuation take advantage of mechanical, electrostatic, piezoelectric, magnetic, thermal, fluidic, acoustic, chemical, biological, or optical principles.

Although optically activated actuators are probably the least developed of all force generating structures, they offer several interesting design features. All-optical circuits and devices have advantages over the conventional electronic components, because they are activated by photons instead of currents and voltages. In many of these designs, the photons provide both the energy and control signal into the system to initiate the desired response. Furthermore, optical systems are free from current losses, resistive heat dissipation, and friction forces that greatly diminish the performance and efficiency of conventional electronic or electro-mechanical systems. The negative effects of current leakage and power loss are greatly amplified, as design engineers strive for product miniaturization through the exploitation of nanotechnology.

Optical actuators can be interfaced with fibre optics and other types of optical wave guide used in developing integrated optical circuits. The opportunity to interface optical actuators directly with fibre optic sensors has enabled a variety of smart devices and engineered structures to be developed. Optical fibre sensors have been designed to measure linear displacement, rotation, force, pressure, sound, temperature, flow, and chemical quantities [2]. Optical sensors and actuators are ideal components for smart structures because they are immune from electromagnetic interference, safe in hazardous or explosive environments, secure, and exhibit low signal attenuation. Systems that directly combine fibre-optic sensors or free-space optical sensors with optical actuators have been termed 'control-by-light' systems (figure C4.3.1). Since the level of light attenuation along the fibre optic is low, the optical sensors and actuators can be located at great distances from the measurement environment in order to ensure electrical isolation.

An optoelectronic control system can be constructed from a variety of lightwave technologies and control strategies. The complexity of the controller design and the control loop is very much application

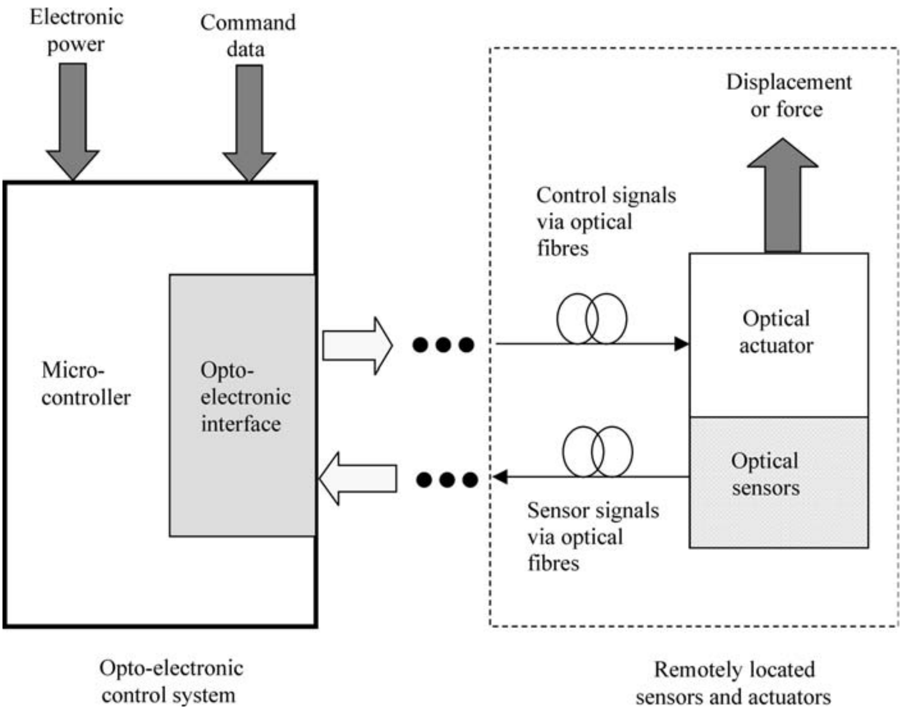

Electronic power
Command data
Micro-controller
Opto-electronic interface
Displacement or force
Control signals via optical fibres
Optical actuator
Optical sensors
Sensor signals via optical fibres
Opto-electronic control system
Remotely located sensors and actuators

Figure C4.3.1. Schematic view of the basic components of a 'control-by-light' system as described by Jones and McKenzie [9].

dependent. It is possible to replace many of the electrical components with optical analogues in order to increase the processing speed or to enhance performance. One possible design for a single-input single-output (SISO) optical controller exploits the functionality of an electro-optic spatial light modulator. A spatial light modulator (SLM) is an optical device that controls the spatial distribution of the intensity, phase, and polarization of transmitted light as a function of electrical signals or a secondary light source [3]. SLMs are the basic information processing elements of nearly all optical systems and provide real time input–output interfaces with peripheral electronic circuitry. Figure C4.3.2 is a block diagram that shows the basic control strategy for regulating the displacement, $\delta(t)$, of a light activated actuator. The electro-optic SLM enables the intensity of the light source to be modified with respect to the error between the reference and feedback values, that is

$$u(t) = K(i_2(t) - i_1(t)) \qquad \text{(C4.3.1)}$$

where K is the fixed system gain and $u(t)$ is the control signal given by light intensity. Most commercially available SLMs have problems with high resolution, large bandwidth, long-term stability, high speed, and low cost. Many of the shortcomings are directly related to the physical limitations of the materials used in the device.

The fundamental and unique characteristics of light activated optical actuators and control systems are explored in this chapter. The more commonly studied light-driven and control-by-light systems that exclusively use off-the-shelf optoelectronic devices, such as photocells and photodiodes, to generate a current to drive an electromagnetic motor directly [4], are not discussed. Here, the primary means of actuation is to project light onto an actuator shell in an effort to generate mechanical deformation that,

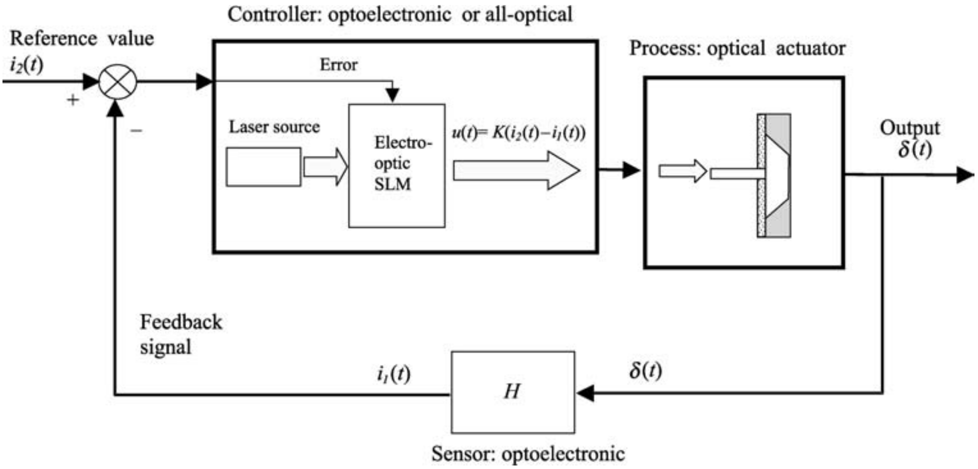


Figure C4.3.2. Block diagram of an optical controller used to regulate the behaviour of a flexible diaphragm optical actuator.

in turn, produces the desired displacement or force. Light is used both to initiate movement and control the actuation mechanism to perform work. The effect of optical actuation and control is illustrated by several innovative system designs.

C4.3.2 Optical actuators

Optical actuators will *indirectly* or *directly* transform light energy into the desired structural displacement. Indirect optical methods exploit the ability of light to generate heat and directly influence the thermal properties of gases, fluids, and solids. On the other hand, direct optical methods use photons to interact with the photosensitive properties of the material used to construct the actuator shell and, thereby, cause mechanical deformation. An example of direct optical microactuation occurs when a light source is used to generate electrostatic forces that move a silicon microcantilever beam [1]. Indirect optical methods often have simpler designs and generate more actuation power than direct optical methods. However, these methods of actuation utilize thermal and phase-transformation properties of fluids and solids, and can be comparatively slow. In contrast, direct optical microactuators are fast but produce small forces. Direct optical microactuators can be designed for a specific application using the proven fabrication techniques of micromachining or semiconductor doping and etching. Direct optical actuators are, therefore, crucial in the development of sophisticated micromachines that are constrained by spatial, low noise, and power requirements. A large number of indirect and direct optical actuators have been reported in the literature. Several of the most promising concepts are now briefly described (table C4.3.1).

C4.3.2.1 Indirect optical actuators

Opto-thermal expansion of fluid

Many indirect optical methods used for mechanical actuation take advantage of the heat generated by the light source to create the desired force or pressure. When a simple gas is heated, it expands

Table C4.3.1. Examples of different optical actuators.

Actuator principle	References
Indirect actuation	
Phase transformatioin of fluids	[5, 1]
Optopneumatic converter	[6, 7, 33, 9, 28, 34, 35, 8]
Phase transformation of solids	[36, 12]
Light propulsion system	[13]
Direct actuation	
Radiation pressure	
Optical micromanipulation	[14, 15, 18, 17]
Solar sails	[19, 20]
Microactuators	
Electrostatic pressure	[1, 21]
Photostrictive effect	[22, 23]
Photothermal vibration	[27]
Photo-induced phase transition material	[25]

according to the ideal gas law

$$P_{\mathrm{r}} V_{1} = n\rho T \tag{C4.3.2}$$

where P_{r} is the gas pressure, V_1 is the gas volume, T is the absolute temperature, ρ is the gas constant (0.0821 litre atm mol^{-1} K^{-1}), and n is the number of moles. The optically actuated silicon diaphragm device shown in figure C4.3.3 exploits this simple principle to deflect a flexible diaphragm in order to perform mechanical work. The cavity is filled with a gas or oil that expands when heated by the light source. As the diaphragm expands under pressure it produces the desired deflection, δ.

The displacement produced by a diaphragm actuator, δ, at the centre from its equilibrium position, is given in [1]

$$P_{\mathrm{r}} = \frac{4a_1 b}{L^2}\sigma_0 \delta + \frac{16a_2 f(\upsilon) b}{L^4}\left(\frac{Y}{1-\upsilon}\right)\delta^3 \tag{C4.3.3}$$

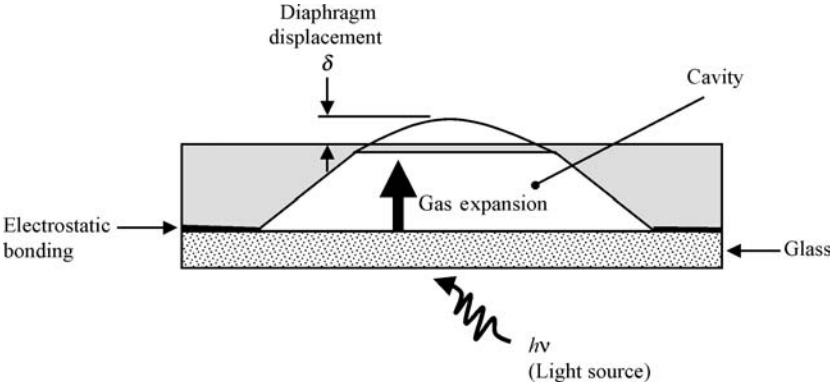


Figure C4.3.3. Schematic view of an optically actuated silicon diaphragm.

where P_r is the applied pressure, L is the length, σ_0 is the residual stress, $Y/(1-\upsilon)$ is the bi-axial modulus, and b is the thickness of the diaphragm. The dimensionless parameters a_1, a_2 and $f(\upsilon)$ depend on the geometry of the diaphragm. For a square diaphragm $f(\upsilon) = 1.075 - 0.292\upsilon$, $a_1 = 3.04$, and $a_2 = 1.37$. Microfabricated flow controllers [1] have speeds of 21 ms in air-flow and 67 ms in oil-flow, and show sensitivities of 304 and 75 Pa mW^{-1}, respectively.

Mizoguchi *et al* [5] used this simple concept to create a micropump that included an array of five closed diaphragm actuated devices called microcells (see figure C4.3.4). Each microcell consisted of a pre-deflected 800 μm × 800 μm square membrane that was micromachined in 0.25 mm^3 of silicon and filled with Freon 113, a liquid with a boiling point of approximately 47.5°C. A carbon-wool absorber was placed inside the cell to convert the incident light from the optic fibre into heat. The microcell exhibited a relatively large deflection, approximately 35 μm, when the cell's contents were heated and the Freon 113 underwent a phase change from liquid to gas. The fluid that is being transported by the pump is fed into a flow channel between the glass plate and deflecting membrane using very small harmonic movements. The harmonic order of the membrane's deflection determines the fluid flow rate and direction. The small quantities of Freon in each cell allowed relatively low optical powers to be used to change the phase of the liquid to gas, giving the large membrane deflections needed to operate the pump. The microcell was fabricated and operated by a laser source with not more than 10 mW. The micropump achieved a head pressure of approximately 30 mmag and flow rate of 30 nl/cycle.

In the late 1980s, a large-scale optopneumatic converter using a closed photoacoustic cell and a flexible membrane nozzle arrangement, figure C4.3.5, was developed at Brunel University [6–8]. The converter was used to operate a double-acting pneumatic cylinder and a standard 21–100 kPa valve actuator in an incremental manner. The basic design was a miniature cylindrical cell (5 mm^3 in volume), which had a thin elastic membrane stretched across one end. An optical fibre rested in a hole on the opposite face of the membrane. A diode laser produced an average of 3.5 mW of optical power that

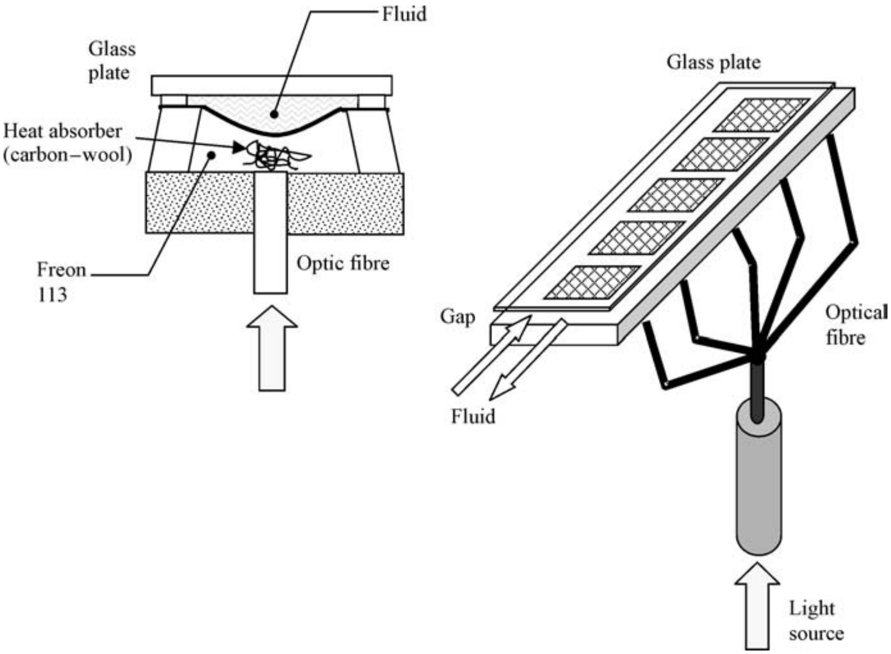


Figure C4.3.4. Cross section of a single microcell in the optically driven micropump. (Adapted from [5].)

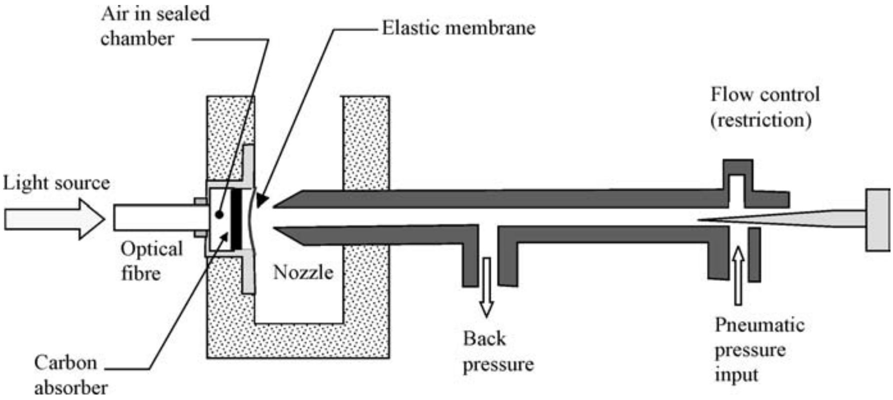


Figure C4.3.5. Schematic diagram of the Brunel converter with an opto-fluidic cell and flapper nozzle valve. (Diagram modified from [6].)

was delivered to the cell. Inside the cell there was a small amount of air and a carbon absorber. The energy transmitted from the light source caused the carbon absorber to heat up, and the transfer of the generated heat to the surrounding air in the cell resulted in an increase in pressure that forced the flexible membrane to expand outward. The membrane movement was detected by a change in the back-pressure of the air jet from the flapper nozzle valve. A differential configuration of two cells was used by Hale *et al* [6] to produce a system that would respond only to changes in the backpressure from two converters. A direct differential pressure of 500 Pa was produced 0.5 s after the laser was switched on, and a maximum of 1400 Pa was produced after the laser was switched on for a longer period of time. The conversion factor is quoted as 400 Pa mW^{-1} [6].

An alternative optical-to-pressure converter was developed by Jones and McKenzie [9], which utilized more recent micromachining techniques to miniaturize the previous design concept and achieve higher response speeds. Light from a 10 mW continuous wave diode laser, delivered by an optic fibre, enters the cell of the converter through the glass cell back. Again, the basic design of the cell contained a fibrous absorber and air. The square cells with integral diaphragms were micromachined with cell volumes of approximately 0.9 mm^3. The converter produced a response time of 21 ms and a conversion factor (pressure/optical power) of 304 Pa mW^{-1} in air, and 67 ms at 81 Pa mW^{-1} in oil.

Phase transformation of solids

Other approaches to indirect optical microactuation exploit the volume expansion and contraction characteristics exhibited by a number of unique materials. These solids experience a discontinuous change in their volume, near the structural phase transformation temperature, that is significantly larger than the linear volume change that occurs due to normal thermal expansion in most materials. Shape memory alloys (SMAs) and some gels, such as polyacrylamide, exhibit reproducible phase transformation effects, and the corresponding phase transformation temperature can be adjusted over a wide range of temperatures making them ideal materials for constructing optically driven microactuators.

SMAs, such as 50/50 Ni–Ti, are a group of metal alloys that can directly transform thermal energy into mechanical work. The shape memory hysteresis effect, figure C4.3.6, is the result of a martensite-to-austenite phase transformation that occurs as the Ni–Ti material is heated, and its subsequent reversal during cooling. The hysteresis effect implies that the temperature at which the material undergoes a phase change during heating is different from the temperature that causes the same material to return to

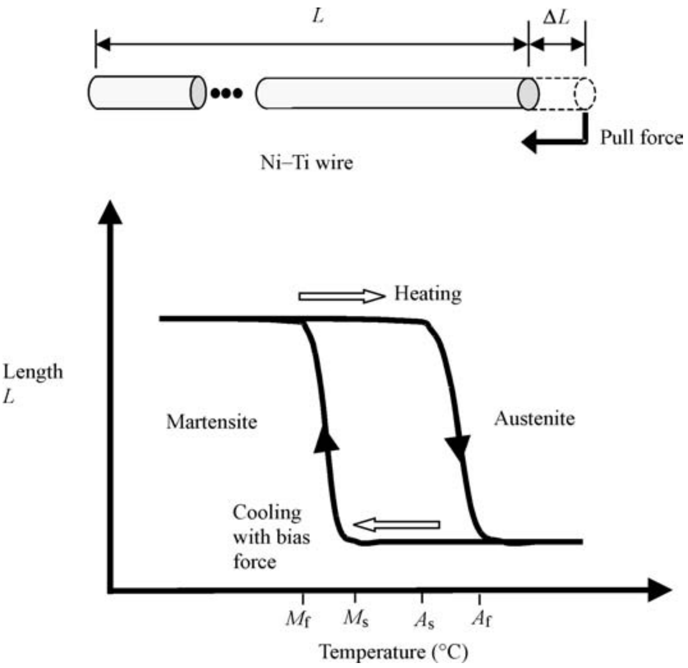


Figure C4.3.6. A schematic of the Ni–Ti wire during contraction and a typical plot of the changes in SMA material properties as the temperature is increased and decreased. In this illustration, A_s is the start of the austenite phase, A_f is the finish of the austenite phase, M_s is the martensite start temperature, and M_f is the martensite finish temperature.

the martensite state during cooling. The hysteresis effect is typically of the order of 20°C for SMA material [10, 11].

The alloy can be formed into a wire or strip at high temperatures when it resides in an austenitic condition. Increasing the temperature of a pre-loaded Ni–Ti wire, originally at ambient temperature, will cause the material to undergo a phase transformation and move the position of the attached load to a distance of approximately 4% of the overall wire length. In essence, the small force created during the contraction period can be used to perform mechanical work [10]. The reduction in the wire length can be recovered by cooling the material back to ambient temperature. The number of times the Ni–Ti material can exhibit the shape memory effect is dependent upon the amount of strain, and consequently, the total distance through which the wire is displaced. The amount of wire deflection is also a function of the initial force applied. The mechanical properties of a typical 200 μm wire are shown in table C4.3.2. The amount of pull-force generated for the applied current is significant for the size of the wire. Thicker wires will generate greater forces but require larger currents and longer cooling time. For example, a 200 μm Ni–Ti wire produces four times more force (~5.8 N) than a 100 μm wire, but takes 5.5 times as long (~2.2 s) to cool down once heating has ceased [11].

Although SMA materials exhibit unique and useful design characteristics, such as large power/weight ratio, small size, cleanness and silent actuation, the successful application of the material has been limited to a small number of actuation devices that require small linear displacements. An example of an optically driven walking machine that employs SMA is shown in figure C4.3.7 [12]. The miniaturized machine consists of two parts: a body made up of SMA and springs, and feet made up of magnets and temperature sensitive ferrites. The feet stick to the carbon steel floor due to magnetic

Table C4.3.2. Mechanical properties of the Ni–Ti wire.

Parameter	Characteristic	Phase
Transformation temperature	90°C	
Density	6.45 g cc^{-1}	
Young's modulus	83 GPa	Austenite
Young's modulus	28–41 GPa	Martensite
Yield strength	195–690 MPa	Austenite
Yield strength	70–140 MPa	Martensite
Transformation strain	Max 8%	For a single cycle
	6%	For 100 cycles
	4%	For 100 000 cycles

Source: [11].

force balance caused by the incident light beam, and the body repeats stretching and shrinking using the deformation of SMAs caused by the switching ON and OFF of the projected light beam.

Certain gels that undergo a phase transformation between a solid and liquid state can also be used as an optically activated microactuator. These gels are mostly fluidic and composed of a tangled network of polymer strands. A balance of forces within the material maintains this state until it is disturbed by very small perturbations introduced by optical, thermal, electrical, or chemical influences. These perturbations cause the material to undergo a phase transformation that can drastically alter the volume of the gel by forcing it to shrink or swell by a factor of several hundred times. For example, the polymer network may lose its elasticity and become compressible as the temperature is lowered. The gel will collapse below a critical temperature because the elasticity becomes zero and compressibility becomes infinite. These changes are discontinuous around the critical temperature and result in large volume changes for an infinitesimal change in temperature.

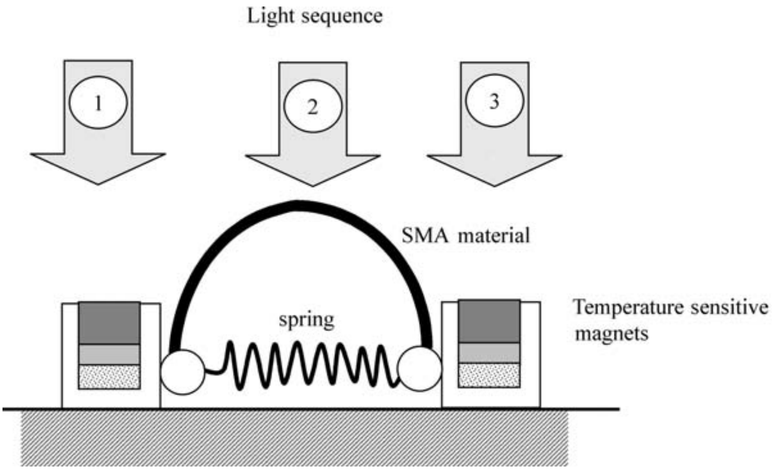


Figure C4.3.7. Basic structure of a light activated walking machine described by Yoshizawa *et al* [12].

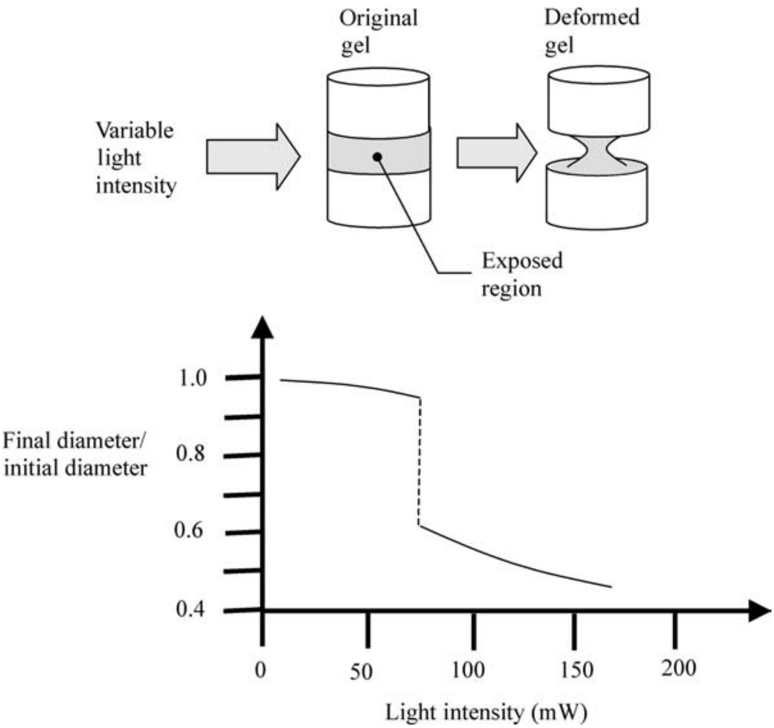


Figure C4.3.8. The diameter of N-isopropylacrylamide/chlorophine copolymer gel as a function of optical power. (Graph adapted from [1].)

Figure C4.3.8 shows the effect of visible light on N-isopropylacrylamide/chlorophine copolymer gel [1]. The ambient temperature of the gel is 31.5°C and the illumination provided by the visible light, with a focused spot size of 20 μm, results in nearly 50% change in the gel's diameter. The actuation speed is several seconds for a gel sample with a diameter of approximately 200 μm. Alternatively, the same actuating material can be illuminated by UV light that initiates an ionization reaction in the gel which creates, in turn, an osmotic pressure that induces swelling. This process is, however, slow when compared to heating the gel directly with white light illumination.

Light propulsion system

Indirect optical actuation has also been proposed as the mechanism for a new and innovative vehicle propulsion system. Recent experiments sponsored by NASA and the US Air Force have demonstrated that a highly polished, reflective object can be propelled into the air by a pulsed IR laser beam originating from the earth's surface [13]. An early functional prototype was constructed from an ordinary aircraft-grade aluminium with a forward covering (or aeroshell), an annular (ring-shaped) cowl, and an aft part consisting of an optic and expansion nozzle. The spin-stabilized object measured 10 cm in diameter and weighed 50 g. By using a 10 kW carbon dioxide laser pulsing 28 times per second, the object has been propelled to altitudes of up to 30 m in approximately 3 s.

The reflective surfaces on the object were used to focus the light beam into a ring, where it heated the surrounding air to a temperature that is nearly five times hotter than the surface of the sun [13], causing the air to expand rapidly and produce the required thrust action. During atmospheric flight, the forward section compresses the air and directs it to the engine inlet. The annular cowl takes the brunt of the thrust. The aft section serves as a parabolic collection mirror that concentrates the IR laser light into an annular focus, while providing another surface against which the hot-air exhaust can press. Myrabo [13] claims that the design offers automatic steering such that if the craft starts to move outside the beam, the thrust inclines and pushes the vehicle back. Although promising, this concept of light propulsion needs further research and development on the optical and control systems in order to realize its potential.

C4.3.2.2 Direct optical actuators

In contrast, direct methods of optical actuation attempt to use the photons in a light stream to induce mechanical deformations within the actuator shell without heating the surrounding gases, liquids, or solids. Although direct optical actuators are fast and energy efficient, they have achieved limited use because they produce relatively small forces and displacements. However, these characteristics and the reduction in resistive heat dissipation and electrical current losses make the direct optical actuators ideal for developing micro-robots, micro-electro-mechanical systems (MEMS), and nanotechnology. Furthermore, many of these optically driven microactuators can be designed for specific applications using proven microfabrication techniques such as laser-material removal, semiconductor doping, and etching.

Optically induced movement by radiation forces

The micromanipulation of particles is, perhaps, the most direct use of light irradiation to both power and control the movement of microscopic machines. Ashkin [14] was the first to suspend and levitate minute particles using the forces created by *light radiation pressure*. A later experiment performed by Ashkin and Dziedzic [15] trapped a 20 μm glass sphere in a vertical laser beam using a 0.25 W TEM_{00} 0.515 μm laser beam. The glass micro-sphere was able to sit in an on-axis position for a period of hours because the transverse forces, caused by the difference in refractive indices of glass and air, pushed the particle towards the region of maximum light intensity, thereby creating a potential well. However, the manoeuvrability of suspended spheres could only be observed by scattering the light of the supporting beam or an ancillary laser with a very low power. Ashkin and Dziedzic [16] further demonstrated that with a small refinement on this simple experiment it was possible to assemble aggregates of two, three and four micro-spheres in mid-air with two laser beams.

The driving force in this micromanipulation technique is the radiation pressure generated by streams of photons striking a surface. An individual photon exhibits a momentum [17] described by

$$M = \frac{hv}{c} \tag{C4.3.4}$$

where h is Planck's constant, v is the optical frequency, and c is the speed of light. The force generated by a single photon is the change in momentum of the photon as it is absorbed or reflected by the surface. If the photon strikes a surface with 100% absorption, then the corresponding force on the structure is

$$F = \frac{\Delta M}{\Delta t} = \frac{hv}{c\Delta t} = \frac{h}{\lambda \Delta t} \tag{C4.3.5}$$

where the optical frequency is given by $v = c/\lambda$ and λ is the wavelength of the light source. In contrast, if the photon strikes a mirror surface with 100% reflectivity, then the force is doubled because the surface is recoiled with enough momentum to stop the photon and send it back, thereby increasing the momentum by a factor of 2.

The total force due to numerous photons striking a mirror surface is estimated as the force–time product generated by a single photon times the number of photons per second as a function of light beam power,

$$F_t = \frac{2h}{\lambda}\left(\frac{P\lambda}{hc}\right) = 2\frac{P}{c} \tag{C4.3.6}$$

where the light beam at power P provides Phc/λ photons per second. Again, if the surface absorbs 100% of the light, then only half the force is generated per photon and the total force is decreased proportionately.

The force generated by light radiation is applicable to both individual particles and ensembles of particles, and has been proposed as the principal technique for optically manipulating small shaped objects for the assembly of micromachines. Higurashi *et al* [18] demonstrated how carefully shaped fluorinated polyimide micro-objects with a cross-sectional radius of 6–7.5 μm can be rotated. In this series of experiments, radiation forces near the focal point are used to position and rotate the micro-object about the laser beam axis as shown in figure C4.3.9. These micro-objects with a low relative refractive index are surrounded by a medium with higher index, and are optically trapped by exerting radiation pressure through their centre openings using a strongly focused trapping laser beam. The pressure is exerted on the inner walls of the object. The micro-objects are both trapped and rotated by radiation pressure when the horizontal cross sections of these objects show rotational symmetry. In addition, the rotation speed versus optical power and the axial position of the laser focal point were investigated for high relative refractive index micro-objects [18]. The rotational speed, with respect to optical power, was found to be in the range of 0.4–0.7 rpm mW^{-1}.

The forces generated by light radiation are also being explored as a means for creating propulsion for space travel [19, 20]. A largely conceptual spacecraft is being developed by the Jet Propulsion Laboratory (JPL) and NASA that uses a solar sail with a very large size, low mass, highly reflective surface that receives a high-powered energy beam from an earth-bound laser. The large size of the sail relative to its mass enables the simple craft to receive sufficient light radiation to propel it forward. The projected photons from the concentrated coherent light source can cause two effects on the sail surface that impact acceleration. First, the photons from the incident light beam collide elastically with the electromagnetic field surrounding the atoms in the sail material and are reflected from the surface. Second, the photons are absorbed by the sail material and generate heat. The amount of usable power transmitted to the sail is constrained by these thermal effects because heating the material reduces its reflective properties. This is a critical design issue because a highly reflective surface will produce a significantly larger force than a light-absorbing surface and, thereby, greater acceleration. However, the temperature on the surface of the sail can be lowered by coating the reverse side with a material that will efficiently radiate most of the generated heat.

During operation, the solar sail must sustain its acceleration in order to reach the high velocities required for travelling great distances in space. In theory, the maximum attainable velocity is limited by the duration of the laser 'efficiently' illuminating the moving target. For long distant travel, the light sail must take full advantage of the coherence in the laser beam. Coherence implies that the energy in the light beam will be reasonably undiminished up to a distance known as the *diffraction distance*. Beyond this point the power from the light source is quickly reduced. Furthermore, the diffraction distance of any laser source is governed by the size of the laser's

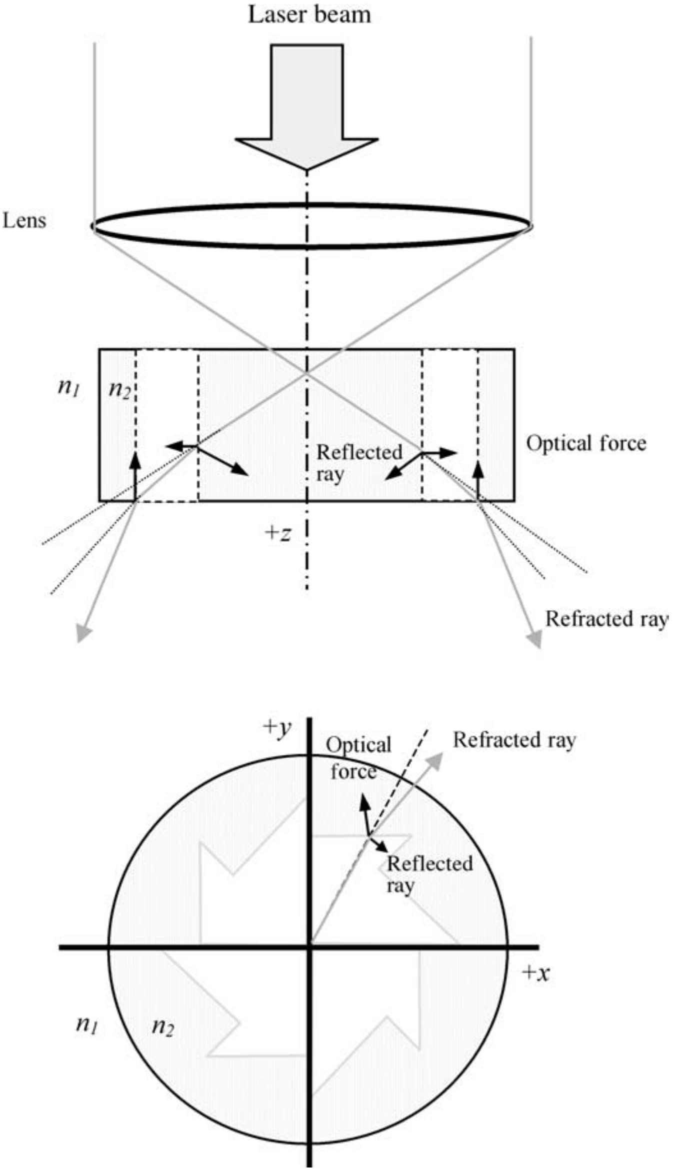

Laser beam
Lens
n_1
n_2
Reflected ray
Optical force
+z
Refracted ray
+y
Optical force
Refracted ray
Reflected ray
+x
n_1
n_2

Figure C4.3.9. Schematic illustrating the origin of optical trapping and optically induced rotation of a low relative refractive index micro-object ($n_1 > n_2$). This experimental set-up was used to observe optically induced rotation of fluorinated polyimide micro-objects. (Illustration modified from [29].)

aperture. A laser system powerful enough to propel a craft, either within or outside the Earth's atmosphere, would probably be constructed from hundreds of smaller lasers arranged in an array. In this case, the effective aperture size of the light source would roughly be the diameter of the entire array. Maximum power will be transferred to the distant sail when an array is packed as densely as possible.

Optically controlled silicon microactuators

Silicon microactuators can be excited directly by an optical light signal using a number of different techniques. One optically controlled silicon microactuator, based on the mechanics of a parallel-plate capacitor as shown in figure C4.3.10, has been developed by Tabib-Azar [1]. The method uses photo-generated electrons to change the *electrostatic pressure* on a thin silicon (Si) cantilever beam mounted on an insulating post overhanging an Au ground plane to form a capacitor given by

$$C \approx \frac{\varepsilon L w}{d} \tag{C4.3.7}$$

where ε is the free-space permittivity, w and L are the width and length of the cantilever beam, respectively, and d is the distance between the cantilever beam and the ground plane.

A potential difference is provided by a power supply V_s that is connected across the capacitor plates through a resistor. The stored charge,

$$Q = CV_s \tag{C4.3.8}$$

causes an electrostatic pressure that deforms the Si cantilever beam and moves it towards the ground plane. When light of sufficient energy is shone onto the metal ground plate, photoelectrons are generated. These photoelectrons migrate through the air gap to the cantilever beam, reducing the charge on the capacitor, and altering the deflection of the beam [1, 21]. It is possible to provide the potential difference needed for the cantilever bias by means of photodiodes or solar cells [9].

If the change in capacitance as the cantilever beam deforms is neglected, then the steady-state deflection δ at the end of the beam can be described in terms of the charge on the capacitor,

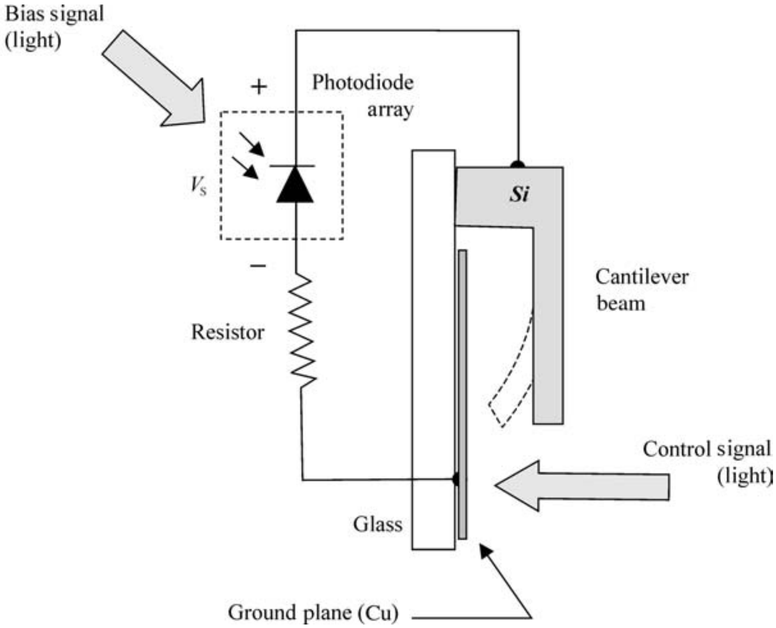


Figure C4.3.10. Proposed design of an all-optically powered and controlled microactuator that produces a displacement based on electrostatic pressure.

Q in equation (C4.3.8), as

$$\delta = \frac{3L^2}{4\varepsilon Y b^3 w^2} Q^2 \tag{C4.3.9}$$

where Y is Young's modulus, and b is the thickness of the cantilever beam. Since this is a nonlinear relationship, the cantilever beam will bend in a stable fashion up to the threshold of spontaneous collapse [1] given by

$$\delta_{\text{threshold}} \approx \frac{6\varepsilon V_2}{Y b^3} \frac{L^4}{8d^2}. \tag{C4.3.10}$$

The fastest time that it can smoothly traverse this distance is approximately equal to the period of the fundamental mode of free vibration.

Tabib-Azar [1] describes a microactuator that uses a $600 \times 50 \times 1\ \mu\text{m}^3$ cantilever beam with a gap of 12 μm. A bias voltage of 6 V and optical power less than $0.1\ \text{mW cm}^{-2}$ was used to move the cantilever 4 μm in approximately 0.1 ms. Continuously charging the capacitor with a current $i \leq i_{\max}/2$, where $i_{\max}$ is determined by the battery circuit, allows light-controlled actuation in either direction. A continuous photon flux, $\Phi < i/\eta$ where η is the quantum efficiency, short circuits the capacitor more slowly than the battery charges it, causing a charge build-up that closes the plates. A photon flux $\Phi > i/\eta$ causes an opposing photocurrent greater than the charging current. The net charge then decreases, and the capacitor plates relax open.

Photostrictive actuators

Optical piezoelectric actuators consisting of PLZT ceramic elements are a new class of actuators that make direct use of both the energy and information being delivered by the incident light. The chemical composition of the ceramic polycrystalline material is given by $\text{Pb}_{(1-y)}\text{La}_y(\text{Zr}_z\text{Ti}_{(1-z)})_{(1-y/4)}\text{O}_3$ where $y = 0.03$ and $z = 0.52$ [22]. The photostrictive effect occurs in the PLZT elements because the material simultaneously exhibits both photovoltaic and piezoelectric properties [23]. In principle, the photovoltaic effect begins when the ferroelectric material is irradiated with light and a large voltage is generated. This photovoltaic effect is different from the phenomenon that occurs in the p–n junction of semiconductors. In this case, it is generated when the electrons are excited by light and move in a certain direction within the ferroelectric crystal due to the spontaneous polarization. Consequently, the piezoelectric effect produces an expansion or contraction of the material when this voltage is generated. The PLZT ceramics exhibit large photostriction characteristics immediately under exposure to light and are, therefore, useful for developing rapid response optical microactuators and photon-driven micromachines.

Morikawa and Nakada [22] produced an optical actuator that had a bimorph structure that was able to generate a displacement of up to several hundred micrometres. The bimorph-type optical actuator, figure C4.3.11, consisted of a pair of PLZT ceramic elements that were adhered in an opposing polarized direction by epoxy adhesive, and common electrodes at both ends. When a 365 nm UV light source irradiated the upper PLZT element in the device, the element stretched in the polarized direction by the photostrictive effect. In contrast, the lower PLZT element was not illuminated by UV light. However, since the lower element is polarized in the opposite direction, it will experience a negative voltage and contract by the piezoelectric effect. The combined effect of the two PLZT elements was that the bimorph-type optical actuator bent downward. Correspondingly, the subsequent illumination of the lower element caused the optical actuator to bend upward.

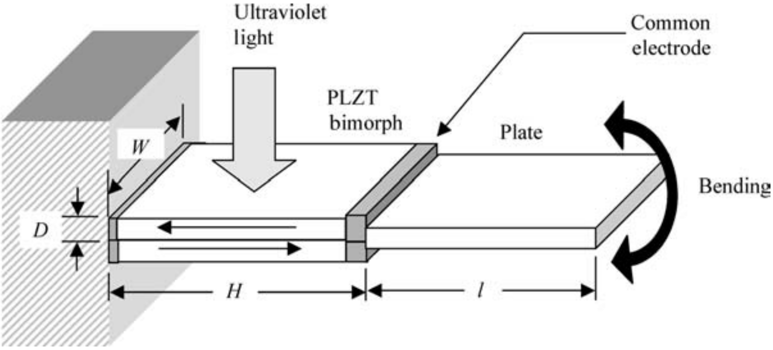


Figure C4.3.11. Structure of a bimorph-type optical actuator and photovoltaic effect of PLZT element. (Modified from [22].)

The displacement of the optical actuator, when only one side of the device is illuminated by UV light of constant intensity I, is the sum of the displacement caused by bending of the PLZT bimorph and the displacement of the plate used for measuring displacement caused by an angle of deflection at the top of the bimorph. The displacement with respect to time can be expressed as

$$\delta(t) = aV_{\mathrm{L}}\left(1 - \mathrm{e}^{-\frac{t}{\tau}}\right), \tag{C4.3.11}$$

where

$$\tau = CR \tag{C4.3.12}$$

and

$$a = \frac{1.5\gamma_{33}}{b}\left(\frac{H}{2} + L\right). \tag{C4.3.13}$$

For equation (C4.3.11), $\delta(t)$ is the displacement of the optical actuator (m), t is the time (s), γ_{33} is the equivalent piezoelectric constant of PLZT element ($\mathrm{m\,V^{-1}}$), H is the distance of the electrodes (m), b is the thickness of the PLZT element (m), L is the length of the plate for measuring displacement (m), V_{L} is the final photo-voltage occurring by the step response (V), τ is the time constant (s), C is the electrical capacitance of the PLZT element (F), and R is the electrical resistance (Ω) of the PLZT element in state of UV irradiation with a given intensity (power/area).

Equation (C4.3.11) shows that the proposed optical actuator has a characteristic first-order lag. The constant a and time constant τ were determined from the dimensions of the PLZT elements. For an equivalent piezoelectric constant of $\gamma_{33} = 370 \times 10^{-12}\,\mathrm{m\,V^{-1}}$ and a UV light source with intensity $I = 1\,\mathrm{kW\,m^{-2}}$, the calculated constants were $a = 337 \times 10^{-6}$ and $\tau = 3.39\,\mathrm{s}$, respectively [22]. The maximum displacement was 183 μm. Fukuda *et al* [24] describe a micro-robotic control system that utilizes a similar bimorph PLZT actuator.

Photo-induced phase transition material

A photo-induced phase transition (PIPT) material will undergo a change in the material's atomic structure when either irradiated by light or exposed to an external electrical field or a temperature gradient. The change in the atomic structure of the PIPT material will cause a direct change in the material specimen's volume and, therefore, a measurable mechanical force and displacement.

Polydiacetylene substituted with alkyl-urethane (PDA-4U3) [25] is an example of a PIPT material that exhibits a reversible phase transition at 125°C between the two stable states. The direction of the transition between the two stable states can be controlled by repeatedly altering the wavelength of the irradiated light (see figure C4.3.12). The response time of PDA-4U3 was observed to be less than 10 ns.

The principal mechanism for mechanical deformation in the PIPT material was the local changes in crystal structure when exposed to light irradiation. If the material is initially in the A state, only the surface region exposed to the light will undergo the change. Since the material in the modified B state has a larger volume in the *c*-axis direction than the original A state, the specimen bends to the side that is opposite to the light irradiated surface (see figure C4.3.13). The observed bending is small, around 2 μm [25]. This implies that the surface thickness contributing to the strain generation in the specimen is much thinner than the overall sample thickness. Furthermore, the deformation

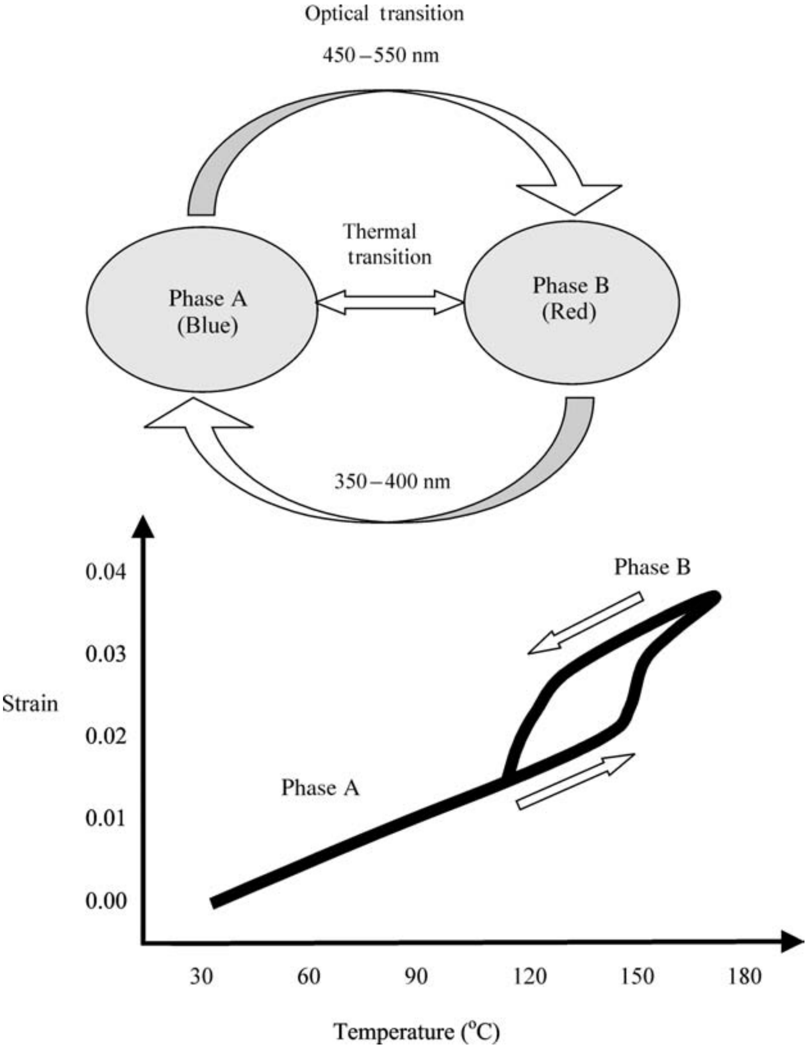

Optical transition
450–550 nm
Thermal transition
Phase A (Blue)
Phase B (Red)
350–400 nm
Phase B
Phase A
Strain
0.04
0.03
0.02
0.01
0.00
30
60
90
120
150
180
Temperature (°C)

Figure C4.3.12. Optical and thermal phase transitions of PDA-4U3 between phases A and B. The light wavelengths shown are for the optical phase transitions. (Graph adapted from [25].)

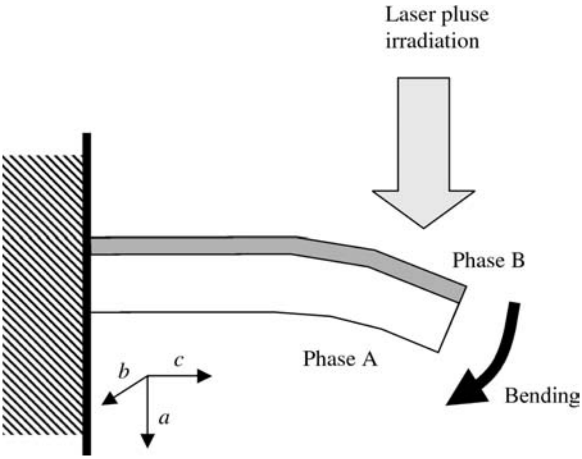


Figure C4.3.13. Mechanism of thin crystal deformation due to light irradiation. Volume expansion corresponding to A to B phase transition at the crystal surface induces a bending moment. (Modified from [25].)

could be sustained without decay for more than 60 min. However, when such an optical actuator is used in a microsystem, the absorption thickness can be designed in such a way that it is comparable to the desired actuator size.

Photothermal vibration

Most of the optical actuators described above produce a static force or displacement that is used to perform mechanical work. In contrast, the following light-driven system generated a dynamic vibrating cantilever beam that can create controlled movement. Inaba *et al* [26] describe how the photothermal vibration of the quartz core of an optical fibre by laser light can be used to construct a vibration-type sensor. The microcantilever beam in this design was the quartz core of the fibre and fabricated by etching the clad layer from the optical fibre tip. The resonance frequency depended largely upon the physical qualities of the cantilever, such as size, density, and Young's modulus. The effect is also partially dependent upon the density of the gas or liquid that surrounds the cantilever, because the *resonance sharpness* of the beam is a function of the viscosity coefficient for the gas or liquid. The resonance frequency for the microcantilever was observed to decrease from 16.69 to 16.59 kHz with an increase in pressure from 1 to 100 Pa, and a reduction in the resonance sharpness with an increase in pressure from 100 Pa to 10 kPa [26].

Based on the concept of photothermal vibration, Otani *et al* [27] proposed a dynamic optical actuator that was driven solely by light. The device is a walking miniature robot constructed from three optical fibres, which represent legs, attached to a base as shown in figure C4.3.14. Each fibre was cut for a bevel and the surface was painted black so that it could absorb light and convert it to heat. The photothermal effect occurred in response to a flashed incident beam, with a constant cycle time, onto one side of the optical fibre leg. The flashing light source produced a stretch vibration on the tip of the fibre that enabled it to operate like a flat spring. The authors experimentally demonstrated that the diameter of the fibre has an influence on the amount of deformation. It was observed that a 10 mm long fibre with a diameter of 250 μm would deform by 30 μm, while a 1000 μm diameter fibre of the same length would deform by as much as 50 μm. Furthermore, Otani *et al* [27] studied the effect of fibre length on the amount of displacement generated. A 1 mm long fibre with a diameter of 250 μm was found to deform 10 μm, while a 15 mm fibre of the same diameter deformed 90 μm.

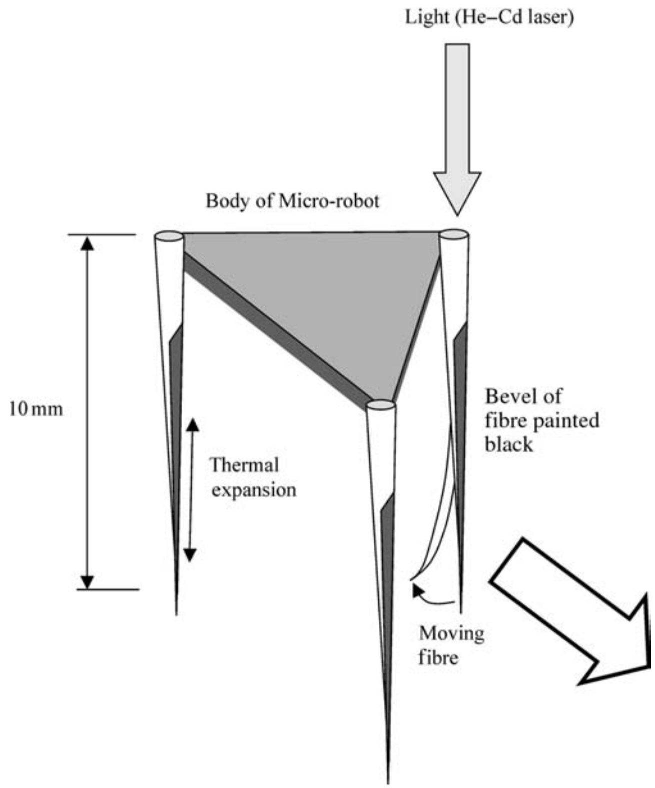


Figure C4.3.14. Optical robot actuated by the photothermal effect on optical fibres as described by Otani *et al* [27].

In the micro-robot design, the optical fibre was bent due to the thermal expansion that occurred when the light was turned on. If the light is turned off, it returns to its original shape. The switching frequency of the 12.1 mW helium cadmium laser (442 nm) was 4 Hz. The light was delivered to an individual optical fibre by a microscopic object lens. An acousto-optic modulator was used to switch the intensity ON and OFF. A mirror and object lens were mounted on a moving stage that followed the movement of the photothermal actuator. The movement of the photothermal actuator also depended on the balance of the device. Otani *et al* [27] adjusted the balance using a small weight on the base of the optical fibre. The size of the optical actuated walking robot was $3 \times 10\,\text{mm}^2$ and it moved 2.3 mm at 25 μm per second using a pulsating light source.

C4.3.3 Optical control

The opportunities presented to a product or process designer by using optically activated actuators and control-by-light systems are now illustrated by considering several innovative examples. The first example describes the optical control of a pneumatic pressure actuator whereas the second summarizes the design and implementation of an opto-hydraulic interface. This discussion is followed by a description of an optical controller used to regulate the behaviour of a piezoelectric actuator in a micro-robot. Finally, the optical implementation of a fuzzy rule-based controller is presented in order to show the broad range of potential applications for this technology.

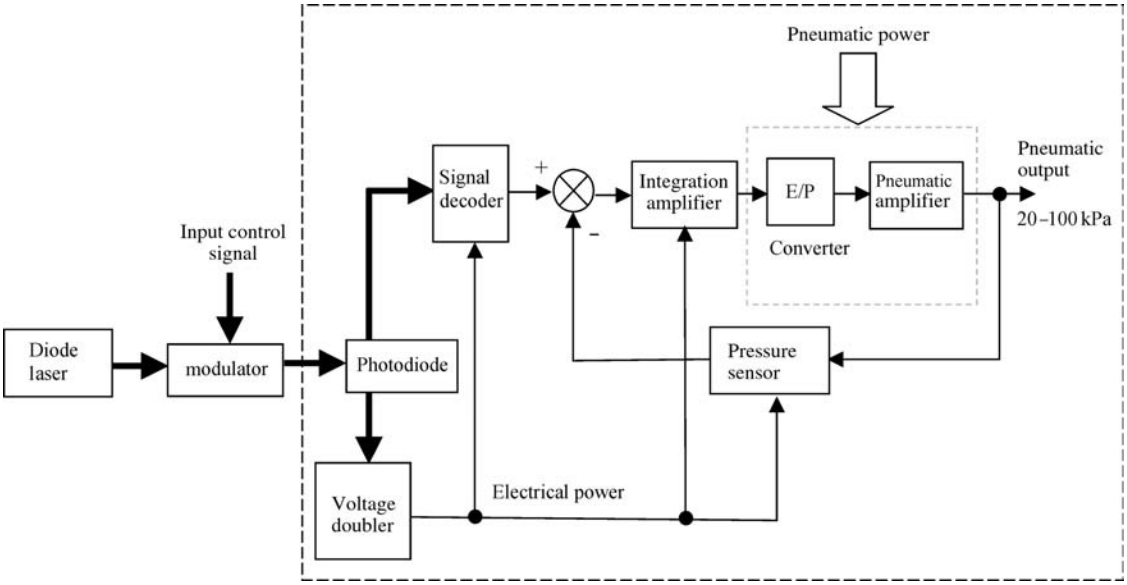


Figure C4.3.15. Configuration of an optically controlled pneumatic pressure actuator. (Diagram modified from [28].)

C4.3.3.1 Optically controlled pneumatic pressure actuator

The control system of the simple optopneumatic pressure actuator was developed by Liu [28] and is shown in figure C4.3.15. The intensity of the laser beam is modulated by an input control signal and transmitted along an optical fibre to the pressure actuator. The optical power in the transmitted signal is then transformed by a photodiode into an electric current that controls a highly sensitive electric-to-pneumatic (E/P) signal converter. The output from this E/P converter is an air pressure signal that is applied to the pneumatic actuator.

Liu's design used a high-impedance moving coil, through a supporting diaphragm to control the back pressure of a pneumatic flapper–nozzle pair. The back pressure was then increased by a volumetric amplifier to give the final actuating pressure. The author demonstrated that a light source with the maximum power at 2 mW and a Si photodetector, with 15% power conversion efficiency, could operate the converter in the pressure range of 20–100 kPa. However, this simple optopneumatic actuator suffered a high nonlinearity in the range of ±10%. These nonlinear characteristics were inherent in the selected photodiode and partly the result of manufacturing errors in fabricating the moving coil in the converter. Additional variations in the output signal were due to light-coupling uncertainties at the fibre connectors and power fluctuations in the light source. This electric-to-pneumatic converter exhibited sensitivity in the range of 119 kPa mA^{-1} [28].

C4.3.3.2 Optical actuation of a hydraulic system

In many product design applications, the optical control signal must be converted to an electronic signal prior to generating the desired response behaviour. However, total immunity to the negative effects of electromagnetic interference can only be achieved when the system actuators are driven directly by the optical signal without an intermediate step for light-to-electrical conversion. A control system developed at United Technologies Research Center in the mid-1980s [29], figure C4.3.16, used an innovative

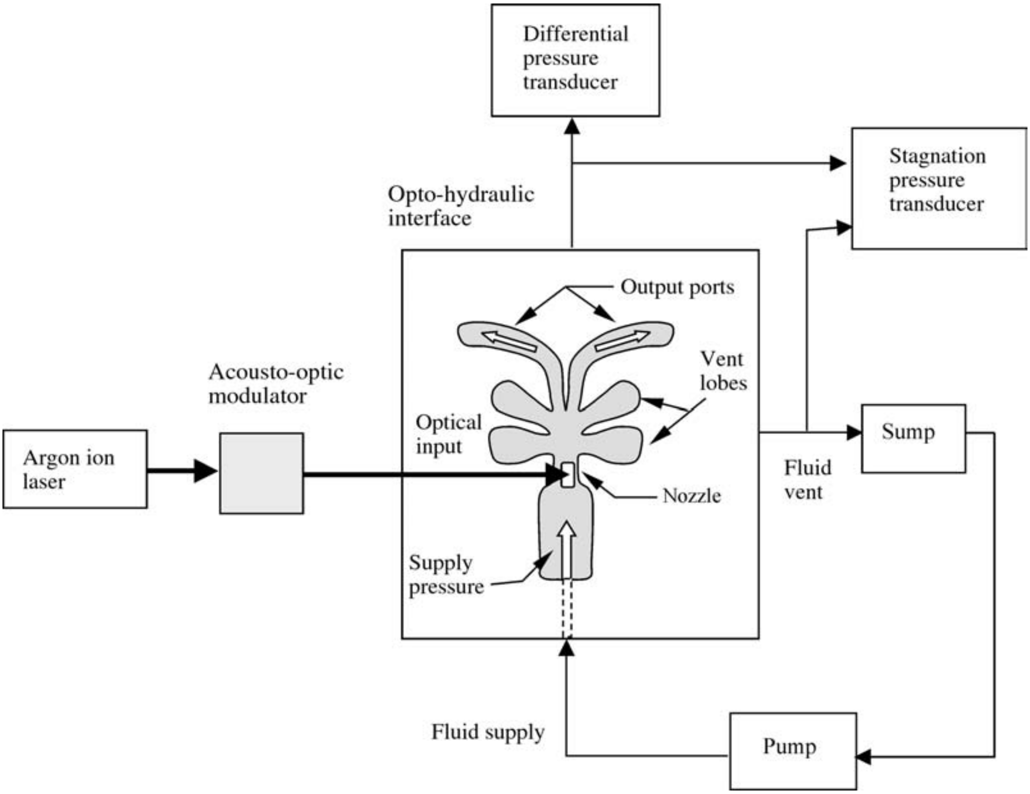


Figure C4.3.16. Block diagram of the opto-hydraulic control system proposed by Hockaday and Waters [29].

optofluidic interface that deflected the fluidic supply jet to perform signal conversion in a hydraulic system. The control computer compared the signal from the hydraulic ram's position sensor with the position command signal, and converted the error into an optical signal that was transmitted to the servo-valve of the hydraulic actuator. At the valve, the optical signal was converted into a fluidic pressure signal and amplified to provide sufficient power to drive the hydraulic servo-valve. Conventional hydraulic hardware was used to operate the ram in the servo-valve.

The optofluidic interface is a modified laminar proportional amplifier (LPA). The basic premise of this device is that optical-to-fluid signal conversion can be achieved by using light to modify the viscosity of the working fluid. In essence, the interface performed the optical signal conversion by indirectly heating the fluid jet. The laminate was sandwiched between two cover plates that had access ports for routing fluid through the passages of the laminate. During operation, a hydraulic pressure source was connected to the supply reservoir so that the fluid accelerated through the nozzle, forming a jet which struck the output ports. The output signal was given as the differential pressure between these ports and was found to be proportional to the angular deflection of the jet. The top cover plate had a window and lens located above the nozzle region providing access for the optical input signal. The bottom cover plate in the nozzle region was fabricated with a graphite–epoxy composite. The composite acted as a black optically absorbent material for converting the optical signal to a thermal gradient across the nozzle region. Conversion of the signal takes place when the optically heated region of the nozzle convectively heats one side of the fluid jet. The resulting thermal gradient produced a viscosity perturbation in the fluid. This perturbation then caused the jet flow to be deflected from the centred position. Thus, the deflected jet struck the output ports unequally producing a differential pressure signal proportional to the optical input.

In the published report [29], the light from the argon-ion 10–30 mW output laser was focused by a microscope objective lens to a 15 μm spot on the absorber, located in an expanding section of the power (supply) jet. The graphite–epoxy absorber transferred the energy from the laser via heat to the sensitive boundary flow layer. This altered the balance of flow between the two output ports. The pressure change was amplified by including additional LPA stages. Using Mil-H-5606 oil, the conversion factor was quoted as 44 Pa mW^{-1} for a 690 kPa supply pressure. The bandwidth for the hydraulic system was reported as 0–145 Hz. By increasing the supply pressure to 1300 kPa, the conversion factor increased to 170 Pa mW^{-1}, and the bandwidth was in the range of 0–170 Hz.

C4.3.3.3 Optical piezoelectric element

Fukuda *et al* [30] demonstrated how the photostrictive effects of a bimorph optical actuator could be used in a servo control system for a variety of micro-robotic applications. The displacement of the bimorph piezoelectric actuator was measured by a noncontact sensor, and the intensity of the UV source irradiation was adjusted by controlling the iris of the optical system. Other types of sensing information such as the temperature at the surface or intensity of illumination could be used for achieving greater precision control. The authors demonstrated that the bimorph element was a multifunctional device because the inherent photorestrictive, pyroelectric, and thermal properties could be used for both actuation and signal communication.

To achieve these goals, Fukuda *et al* [24] developed a method for exposing both sides of the bimorph element to light in order to increase response speed and reduce hysteresis, thereby improving the photoresponse characteristics of the actuator. An optical servo system was constructed for implementing the method and subjected to optical servo tests with PI control. Based on the measured photoresponse of the bimorph PLZT element, the transfer function is expressed as

$$G_{\mathrm{p}2}(s) = \frac{3\Gamma^2}{8L}\left[\frac{\alpha\delta}{(s+\tau_1)(s+\tau_2)} + \frac{a(1-T)}{2L\kappa}\left(\frac{sP_{\mathrm{s}}\delta}{s+\tau_2}+\gamma\right)\frac{1}{s+\frac{a\beta}{2L}}\right] \qquad \text{(C4.3.14)}$$

where Γ (2.1×10^{-2} m) is the length of the optical piezoelectric element, L (2×10^{-4} m) is the thickness, α (4.2×10^{-12} C m^2 W^{-1}) is the conversion coefficient for converting input light energy into electric current, δ (1.5×10^4 C^{-1}) is the strain conversion coefficient, τ_1 (1.5 s^{-1}) and τ_2 (0.5 s^{-1}) are the natural discharge constants, a (2×10^{-5} m^2 s^{-1}) is the thermal diffusivity, T (0.95) is the constant representing the thermal insulation of the epoxy layer as determined through experiments, κ (10 W m^{-1} K^{-1}) is the thermal conductivity, P_{s} (7×10^{-9} C K^{-1}) is the pyroelectricity conversion constant, γ (1×10^{-5} K^{-1}) is the thermal coefficient of the optical piezoelectric element, and $\beta = h/\kappa$ (1.5 m^{-1}) is the thermal radiation constant complying with Newton's law of cooling. Typical values for the parameters are shown in the parenthesis. The input to the system given by equation (C4.3.14) is the thermal flux incident on the PLZT element and the output is the displacement of the front end of the bimorph material. The change in the temperature is approximated by a first-order delay for practical use or simplification of the photoresponse model.

Since a pulse width modulated (PWM) light source is used to activate the bimorph PLZT actuator, the transfer function of the PWM UV irradiator must also be determined. In this application, the pulsed light source was created by rotating a slotted disk, with slits at equal 90° intervals, through 45° rotations by a stepper motor. The interval between two consecutive 45° movements determined the ratio of irradiation time to interruption time, in other words the duty ratio. From experiments, Fukuda *et al* [24] determined the PWM irradiator transfer function to be

$$G_{\mathrm{p1}}(s) = \frac{650}{s+100}\mathrm{e}^{-0.025s} \tag{C4.3.15}$$

where the input to the transfer block is the duty ratio and the output is a thermal flux resulting from the UV irradiation.

Figure C4.3.17 is a block diagram of the servo control system for the optically activated bimorph PLZT actuator. The optical servo system comprises of a position-integral (PI) controller, a signal converter that transforms the continuous control signal into a duty ratio, a PWM UV irradiator, and the bimorph PLZT actuator. The PI controller generates the control signal $u(t)$ which is delivered to the rule based converter for PWM. This signal is a function of the error $e(t)$ between the actual displacement of the front end of the bimorph PLZT element, $y(t)$, and the target displacement, $r(t)$. The rule based converter receives the control signal and transforms it according to

$$d(t) = \begin{cases} 85 < u(t) & 85 \\ 15 \leq u(t) \leq 85 & u(t) \\ -15 < u(t) < 15 & 0 \\ -85 \leq u(t) \leq -15 & u(t) \\ u(t) < -85 & -85 \end{cases} \tag{C4.3.16}$$

where the duty ratio $d(t)$ delivered to the PWM UV irradiator ranges from 15 to 85%. The basic procedure is such that the left side of the PLZT element is exposed when $d(t) > 15$, the right side is exposed when $d(t) < -15$, and neither side is exposed when $d(t) = 0$. The thermal flux created by the PWM UV irradiator, $w(t)$, is then transferred to the bimorph actuator.

Experiments were performed by Fukuda *et al* [24] to test the optical servo system using the following PI controller

$$G_{\mathrm{PI}}(s) = K_{\mathrm{p}}\left(1 + \frac{1}{T_1 s}\right) \tag{C4.3.17}$$

where K_{p} was the comparison gain and T_1 was an integrated time. The actuator's response was examined when the target displacement was 30 mm and the parameters of the PI controller were $K_{\mathrm{p}} = 3$ and

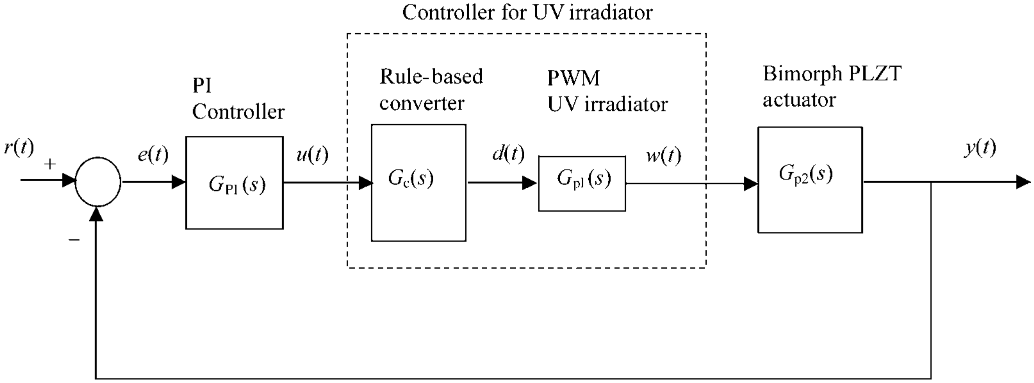


Figure C4.3.17. Block diagram of an optical servo control system. (Adapted from [24]).

$T_1 = 1$, $K_p = 3$ and $T_1 = 10$, and $K_p = 5$ and $T_1 = 10$. The actuator output $y(t)$ was measured using a noncontact eddy current displacement meter. The intensity of the UV light was a constant 170 mW cm^{-2}. In addition, the sampling time was 0.4 s and the measuring time was 30 s. The response [24] showed an overshoot at the rise edge that lasts approximately 2 s before the output returned to its target, $r(t)$.

To demonstrate the practicality of the proposed control-by-light system, Fukuda *et al* [30] described two interesting applications of the bimorph optical actuator. The first example was an optical gripper that used two bimorph actuators to activate gripper movements. UV light was delivered to the gripper device by an optical fibre and mirror subassembly. The measured displacement of the optical gripper tip was approximately 100 μm. Since the gripper design had strong anti-noise characteristics, it has a potential in microsurgical applications. The second illustrative example presented by Fukuda *et al* [30] was an optically controlled micro-robot that used two bimorph piezoelectric actuators to create walking movement. Again, a UV source delivered by optic fibres was used to irradiate the optical piezoelectric element. The micro-robot motion was achieved by controlling the UV beam irradiation pattern to the microactuator and produced a maximum speed of 16 μm s^{-1}.

C4.3.3.4 SISO rule-based optical controller

A large number of consumer products and industrial processes use fuzzy logic techniques to control electro-mechanical components. These rule based controllers often fail to perform in real time because of the extensive signal processing, and algorithmic computations required to arrive at a satisfactory conclusion. Itoh *et al* [31] describe an optoelectronic controller that performs fuzzy logic operations in real time by utilizing a scanning beam fuzzy inference architecture. The beam-scanning laser diode (BSLD) [31] of the SISO system receives a current signal as an input $i_1(t)$ from a sensor monitoring the response of a physical process and a reference input given by current signal, $i_2(t)$. The light beam emitted by the BSLD diverges by an angle θ that is a function of the current error, $(i_2(t) - i_1(t))$. The angle θ represents the basic *premise* for each rule in the fuzzy inference engine. The proposed architecture, figure C4.3.18, uses a product–sum gravity method with Gaussian membership functions instead of the conventional min–max gravity method with triangular membership functions [32].

The deflected beam strikes an array of photodetectors (PDs) where each PD represents a membership function for the angle θ. The central PD represents angles close to zero, and the PDs on the edge of the array represent large angles, either positive large (PL) or negative large (NL). Each constituent PD produces a current whose magnitude is proportional to the match between θ and the PD location. In addition, each PD in the array drives a separate beam scanning laser diode (BSLD). When activated, this BLSD emits light with an intensity that is proportional to the driving current generated by the activated PD. The resultant beams have Gaussian profiles such that the neighbouring BLSDs project onto a common position-sensitive detector (PSD). The output current of the PSD is proportional to the centre of gravity (CoG) of the incident beams.

The main advantage of the optical inference engine proposed by Itoh *et al* [31] is its simplicity and modularity for extending the basic structure to applications that require multiple inputs. For example, several SISO controllers can be placed on top of each other to control several systems simultaneously. However, the basic controller design has several limitations. First, the beam-scanning laser diode and other signal deflection methods generate beams with Gaussian profiles. The reshaping of the beam to represent other types of membership function is not a trivial task. Solutions to this problem include the use of an amplitude mask, reshaping the beam in the Fourier domain with a phase-only filter, or placing an amplitude-coded mask in the Fourier plane. A second limitation to this optical implementation of a fuzzy logic controller is that this optic circuit requires optical–electronic–optical conversions. As a result of these signal conversions, the original information can easily become distorted and lead to control error.

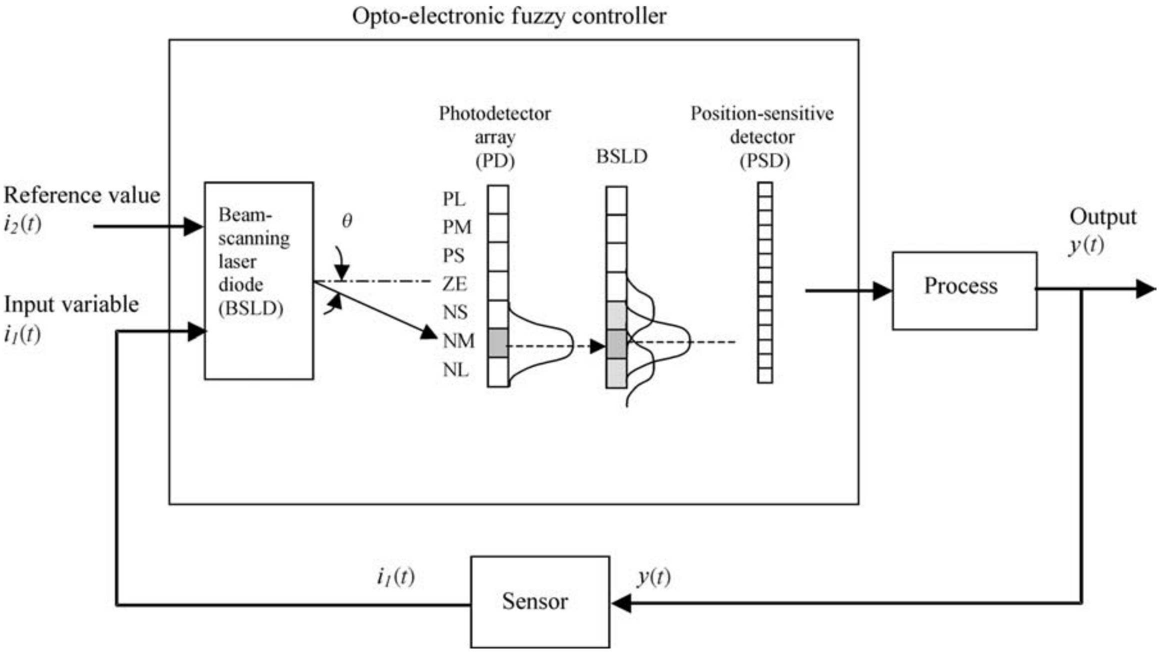
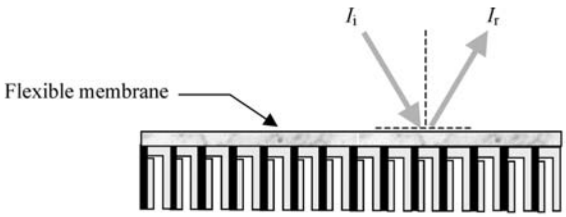


Figure C4.3.18. The optoelectronic implementation of a SISO fuzzy logic controller. In the above illustration the fuzzy membership functions are: PL positive large, PM positive medium, PS positive small, ZE zero, NS negative small, NM negative medium, and NL negative large. (Adapted from [31].)

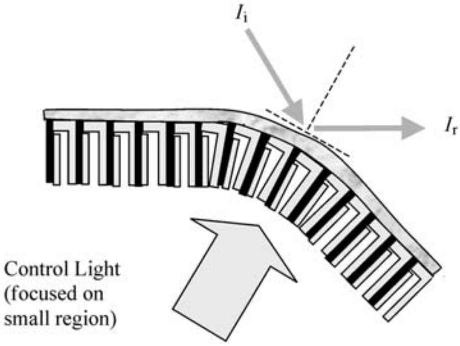
C4.3.4 Conclusions

This chapter has summarized some of the unique optical actuators and control-by-light systems that have been developed and reported in the published literature over the past two decades. Many of the optical actuators provide robust, low cost solutions that enable high precision control and rapid signal processing. All optical systems have advantages over conventional electronic components because they are free from electric current losses, resistive heat dissipation, and friction forces that greatly diminish the performance and efficiency of conventional electro-mechanical systems. The negative effects of current leakage and power loss are greatly amplified as design engineers look toward product miniaturization and the exploitation of recent advances in nanotechnology. The ease in which optical actuators can be interfaced with fibre optics and other types of optical waveguide provides an added incentive to re-evaluate the notion of practical integrated optical circuits.

The opportunity to interface optical actuators directly with fibre optic sensors has enabled a variety of innovative devices and engineered systems to be developed. These *smart structures* are able to modify physical characteristics in response to a light stimulus. These physical characteristics may influence the generated force, structure's elasticity, geometric dimensions, or overall shape. Figure C4.3.19 is an illustration of a smart, reconfigurable mirror with a flexible membrane that encloses numerous embedded light activated silicon microactuators. The discrete cantilever shells can be distributed throughout the structure enabling a focused light source to initiate local changes in shape. Only the microcantilever shells that are directly exposed to the light source will be deflected by a small pre-determined amount δ and, thereby, contribute to the overall curvature change in the mirror. Another example is a smart fabric that can change its elasticity when exposed to an appropriate wavelength or intensity of light.



(*a*) Original shape of a smart structure with numerous microactuators.



(*b*) Smart structure deforms only in the region that is exposed to high intensity light.

Figure C4.3.19. A schematic drawing of a reconfigurable mirror constructed from numerous optically controlled silicon microactuators.

Future developments in optical actuators will arise from the introduction of new light sensitive materials, advances in microfabrication techniques, and the aggressive exploitation of the optical power available in light. Phase transformation materials such as hydrolysed gels and light sensitive proteins hold particular promise in constructing reconfigurable microstructures and machines. These materials allow physical changes or transitions at the molecular level without the need to construct large numbers of discrete complex multi-material devices or actuator shells. Advances in precision micromachining and material deposition methods enable engineers to explore new design concepts. Furthermore, designers will exploit the fundamental physics of light and light interaction with molecules rather than developing microactuators that mimic traditional mechanical linkages, rotors, and pumps [1]. Finally, the energy in light is one of the easiest forms of energy to shape and transmit through free space. Both the power to drive a system and the signal to control its behaviour can be delivered to a precise location with very little distortion or noise. The ease and accuracy with which light energy can be transmitted to individual actuators may lead to innovative machine designs that allow the time-sharing of light beams over great distances.

References

[1] Tabib-Azar M 1998 *Microactuators: Electrical, Magnetic, Thermal, Optical, Mechanical, Chemical, and Smart Structures* (Norwell, MA: Kluwer)
[2] Allard F C 1990 *Fibre Optics Handbook: For Engineers and Scientists* (New York: McGraw-Hill)
[3] Saleh B E A and Teich M C 1991 *Fundamentals of Photonics* (New York: Wiley)

[4] Landry M J, Rupert J W and Mittas A 1991 Power-by-light systems and their components: an evaluation *Appl. Opt.* **30** 1052–1061

[5] Mizoguchi H, Ando M, Mizuno T, Takagi T and Nakajima N 1992 Design and fabrication of light driven pump *Micro Electro Mech. Syst. '92* pp 31–36

[6] Hale K F, Clark C, Duggan R F and Jones B E 1988 High-sensitivity optopneumatic converter *IEE Proc.* D **135** 348–352

[7] Hale K F, Clark C, Duggan R F and Jones B E 1990 Incremental control of a valve actuator employing optopneumatic conversion *Sensors Actuators* **A21–A23** 207–210

[8] McKenzie J S, Hale K F and Jones B E 1995 Optical actuators *Advances in Actuators* ed A P Dorey and J H Moore (Bristol: Institute of Physics Publishing) pp 82–111

[9] Jones B E and McKenzie J S 1993 A review of optical actuators and the impact of micromachining *Sensors Actuators* A **37–38** 203–207

[10] Conrad J M and Mills J W 1999 *Stiquito for Beginers: an Introduction to Robotics* (Los Alamitos, CA: IEEE Computer Society)

[11] Gilbertson R G 1993 *Muscle Wires Project Book* (San Rafael, CA: Mondo-Tronics)

[12] Yoshizawa T, Hayashi D, Yamamoto M and Otani Y 2001 A walking machine driven by a light beam *Optomechatronic Systems II (Proc. SPIE vol 4564)* ed H-Y Cho pp 229–236

[13] Myrabo L N 1999 Highways of light *Sci. Am.* **280** 2 88–89

[14] Ashkin A 1970 Acceleration and trapping of particles by radiation pressure *Phys. Rev. Lett.* **24** 156–159

[15] Ashkin A and Dziedzic J M 1971 Optical levitation by radiation pressure *Appl. Phys. Lett.* **19** 283–285

[16] Ashkin A and Dziedzic J M 1980 Observation of light scattering using optical levitation *Appl. Phys.* **19** 660–668

[17] Katagiri Y 2001 Optical micromachines for photonic networks *Optomechatronic Systems II (Proc. SPIE vol 4564)* ed H-Y Cho pp 152–163

[18] Higurashi E, Ohguchi O, Tamamura T, Ukita H and Sawada R 1997 Optically induced rotation of dissymmetrically shaped fluorinated polyimide micro-objects in optical traps *J. Appl. Phys.* **82** 6 2773–2779

[19] Chmielewski A B, Moore C and Howard R 2000 *The Gossamer Initiative: 2000 IEEE Aerospace Conf. Proc.* **7** 429–438

[20] Whites K W and Knowles T R 2001 Calculating radiation force on carbon fibre Gossamer space sailcraft *2001 IEEE Int. Symp. Antennas Propagation Soc.* vol 3 pp 326–329

[21] Tabib-Azar M and Leane J S 1990 Direct optical control for a silicon microactuator *Sensors Actuators* **A21–A23** 229–235

[22] Morikawa Y and Nakada T 1997 Position control of PLZT Bimorph-type optical actuator by on–off control *IECON Proc.* Part 3 1403–1408

[23] Uchino K 1990 Photorestrictive actuator *IEEE Ultrason. Symp.* pp 721–723

[24] Fukuda T, Hattori S, Arai F and Nakamura H 1995 Performance improvement of optical actuator by double side irradiation *IEEE Trans. Ind. Electron.* **42** 455–461

[25] Ikehara T, Tanaka M, Shimada S and Matsuda H 2001 Optically-driven actuator using photo-induced phase-transition material *14th IEEE Int. Conf. on Micro Electro Mech. Syst.* pp 256–259

[26] Inaba S, Kumazaki H and Hane K 1995 Photothermal vibration of fibre core for vibration-type sensor *Japan. J. Appl. Phys.* **34** 2018–2021

[27] Otani Y, Matsuba Y and Yoshizawa T 2001 Photothermal actuator composed of optical fibres *Optomechatronic Systems II (Proc. SPIE vol 4564)* ed H-Y Cho pp 216–219

[28] Liu K 1991 Power budget considerations for optically activated conventional sensors and actuators *IEEE Trans. Instrum. Meas.* **40** 25–27

[29] Hockaday B D and Waters J P 1990 Direct optical-to-mechanical actuation *Appl. Opt.* **29** 4629–4632

[30] Fukuda T, Hattori S, Arai F, Matsuura H, Hiramatsu T, Ikeda Y and Maekkawa A 1993 Characteristics of optical actuator-servomechanisms using bimorph optical piezoelectric actuator *IEEE Conf. Robotics Automation* **2** 618–623

[31] Itoh H, Yamada T, Mukai S, Watanabe M and Brandl D 1997 Optoelectronic implementation of real-time control of an inverted pendulum by fuzzy-logic-control units based on a light-emitting-diode array and a position-sensing device *Appl. Opt.* **36** 4 808–812

[32] Kosko B 1997 *Fuzzy Engineering* (Upper Saddle River, NJ: Prentice-Hall)

[33] Johnson M 1986 Optical-actuator frequency-coded pressure sensor *Opt. Lett.* **11** 9 587–589

[34] Liu K and Jones B E 1989 Pressure sensors and actuators incorporating optical fibre links *Sensors Actuators* **17** 501–507

[35] McKenzie J S and Clark C 1992 Highly sensitive micromachined optical-to-fluid pressure converter for use in an optical actuation scheme *J. Micromech. Microeng.* **2** 245–249

[36] Suzuki A and Tanaka T 1990 Phase transition in polymer gels induced by visible light *Nature* **346** 345–347

C4.4
Optical to electrical energy conversion: solar cells

Tom Markvart

C4.4.1 Introduction

Photovoltaics is about to celebrate 50 years of its existence in modern era. Photochemical reactions that produce energy from sunlight have been known for over 150 years [1] and photosensitive semiconductor devices were discovered in the second half of the 19th century [2, 3]. It was not, however, until the 1950s that the first practical solar cell was developed at Bell Laboratories in New Jersey [4].

Solar cells soon found applications in supplying electrical power to the first satellites. Terrestrial systems followed in the 1970s. They were what we now call ‘remote industrial’ or ‘professional’ applications in isolated locations, powering radio repeater stations, control and measurement devices, or the cathodic protection equipment for oil and gas lines. Numerous photovoltaic systems have since been installed to provide electricity to the large number of people who do not have (nor, in the foreseeable future, are likely to have) access to mains electricity.

Since the time of the oil crises in the 1970s, solar electricity is being increasingly mentioned as an alternative to conventional electricity supplies which, in most countries, continue to depend on fossil fuels. The first PV power plants were build in the USA and Europe during the 1980s. The most recent and arguably most exciting application came during the last decade of the 20th century: solar cells integrated into the roofs and facades of buildings, as a new, distributed, form of power generation (see table C4.4.1).

Solar cells capable of generating almost 400 MW were produced in 2001, providing electricity for applications ranging from milliwatts to megawatts (figure C4.4.1). This chapter aims to give an overview of this rapidly growing industry, with an outline of the device as well as system aspects of the field. Solar cell operation is described in section C4.4.2 which also examines fundamental constraints on the solar cell efficiency. Section C4.4.3 gives a review of the current photovoltaic technologies, with a brief look at crystalline silicon, thin films and dye-sensitized solar cells. Photovoltaic systems are discussed in section C4.4.4, with an overview of the design and principal features of stand-alone and grid connected systems.

C4.4.2 Principles of solar cell operation

From an engineering point of view, the solar cell can be considered as a semiconductor diode connected in parallel with a current source of the photogenerated current I_l (figure C4.4.2). This argument yields the Shockley ideal solar cell equation

$$I = I_l - I_0\left(e^{\frac{qV}{kT}} - 1\right) \tag{C4.4.1}$$

Table C4.4.1. The growth of photovoltaic applications.

1950s	First modern solar cells
1960s	Solar cells become main source of electric power for satellites
1970s	First remote industrial applications on the ground
1980s	Solar cells used in rural electrification and water pumping. First grid connected systems
1990s	Major expansion in building-integrated systems

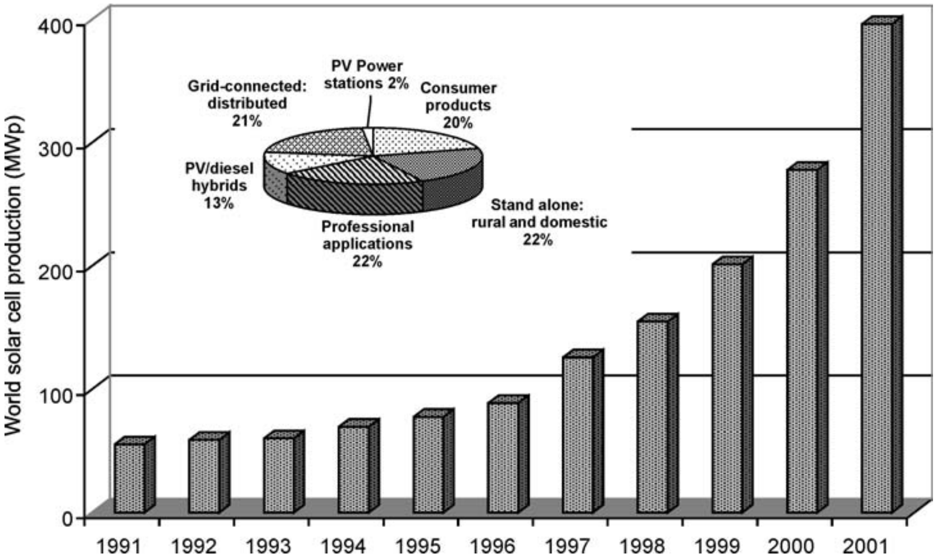


Figure C4.4.1. The world annual production of solar cells, and their principal applications. (Courtesy P Maycock, PV News. Data for applications based on 1997.)

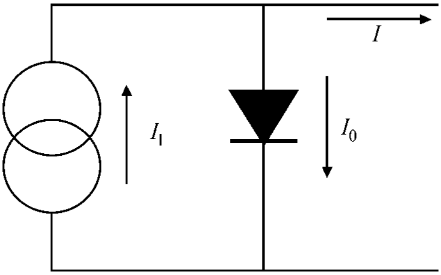


Figure C4.4.2. The equivalent circuit of an ideal solar cell.

where I_0 is the dark saturation current of the diode, q is the electron charge, V is the voltage at the terminals of the solar cell, k is the Boltzmann constant and T is the absolute temperature. The solar cell characteristic is shown in figure C4.4.3, depicting also the relevant parameters I_{sc}, V_{oc} and P_{max}. The short-circuit current I_{sc} produced by this ideal solar cell is equal to the light generated current I_l.

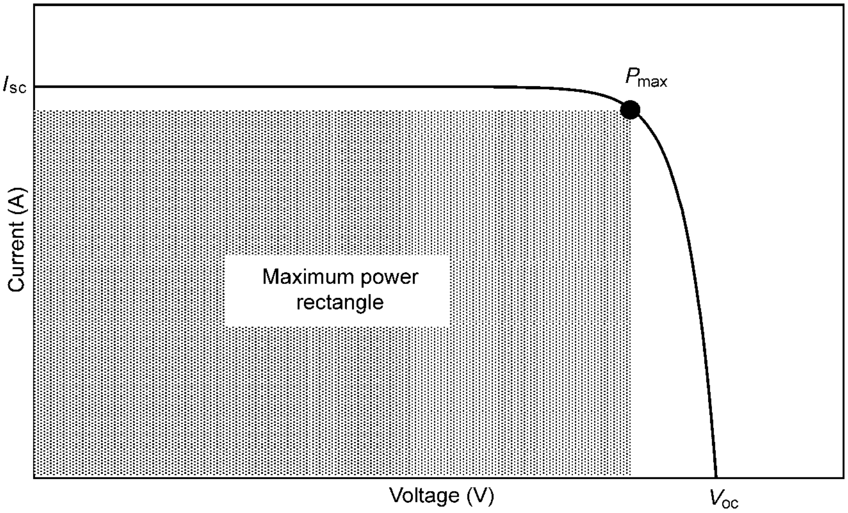


Figure C4.4.3. The *IV* characteristic of a solar cell. The power generated at the maximum power point P_{max} is equal to the area of the maximum power rectangle.

The open circuit voltage V_{oc} is given by

$$V_{oc} = \frac{kT}{q} \ln\left(1 + \frac{I_l}{I_0}\right). \tag{C4.4.2}$$

It is customary to define the fill factor (ff) which completes the calculation of the power produced by the cell at the maximum power point P_{max}:

$$\text{ff} = \frac{P_{max}}{I_{sc} V_{oc}}. \tag{C4.4.3}$$

To describe practical devices, a diode ideality factor n is sometimes introduced as a coefficient of the thermal voltage kT in the exponent of equation (C4.4.1). A second diode with $n = 2$ can also be added, as well as series and parallel resistances to improve agreement with the measured data.

Constraints to solar cell efficiency imposed by fundamental physical laws have been reviewed on a number of occasions [5–10]. The best known of these, the Shockley–Queisser detailed balance limit is briefly reviewed below. These limits reflect the fact that the solar cell is a quantum energy converter, which is not subject to some of the usual constraints which apply to the conversion of heat to mechanical work. Notwithstanding, we shall see that the Carnot factor does appear in a subtle way.

Before the efficiency can be discussed, a common standard of solar radiation has to be agreed upon. Workers in the area of terrestrial solar cells and systems use the most frequently accepted standard, which represents a good compromise between the convenience and observation: solar spectrum at air mass (AM) 1.5, normalized to total energy flux density of 1 kW m^{-2} [11]. The value of 1 kW m^2 for the total irradiance is particularly convenient for the system design (see section C4.4.4) and, together with the cell temperature of 25°C, corresponds to the usual conditions for the calibration of terrestrial solar cells and modules. The rated solar cell output at the maximum power point under these standard test conditions is usually called 'peak watts', and denoted by W$_p$.

Space solar cells and systems operate under extraterrestrial (AM0) radiation, and this spectrum is therefore used for their calibration [12]. The total irradiance, equal to the average solar irradiance outside the earth's atmosphere, is called the solar constant S. The commonly accepted value of S is now $1.367\,\mathrm{kW\,m^{-2}}$.

Theoretical calculations are often based on a third definition of solar radiation: the radiation of a black body which agrees with the observed AM0 irradiance, giving the appropriate temperature of solar radiation $T_s = 5767\,\mathrm{K}$ [13]. To complicate matters further, some calculations use a less accurate but more convenient value of $T_s = 6000\,\mathrm{K}$. A geometrical factor f_ω is introduced to allow for the size of the solar disc as perceived from the earth:

$$f_\omega = \left(\frac{R_S}{R_{SE}}\right)^2 = \frac{\omega_S}{\pi} \tag{C4.4.4}$$

where R_S is the radius of the sun, R_{SE} is the mean distance between the sun and the earth, and $\omega_S = 6.85 \times 10^{-5}$ sr is the solid angle subtended by the sun. The spectra which correspond to these different types of solar radiation are shown in figure C4.4.4.

In the calculation of their detailed balance limit, Shockley and Queisser [9] use the black body photon flux on earth received by a planar surface of unit area in a frequency interval $\delta\nu$, equal to

$$\delta\phi = \frac{2\pi}{c^2} f_\omega \frac{\nu^2 \delta\nu}{e^{\frac{h\nu}{kT_S}} - 1} \tag{C4.4.5}$$

where h is the Planck constant and c is the speed of light. An ideal solar cell made from a semiconductor with bandgap E_g absorbs all photons with frequency $\nu > \nu_g = E_g/h$ and each such photon gives rise to an electron in the external circuit. The short circuit current density is then

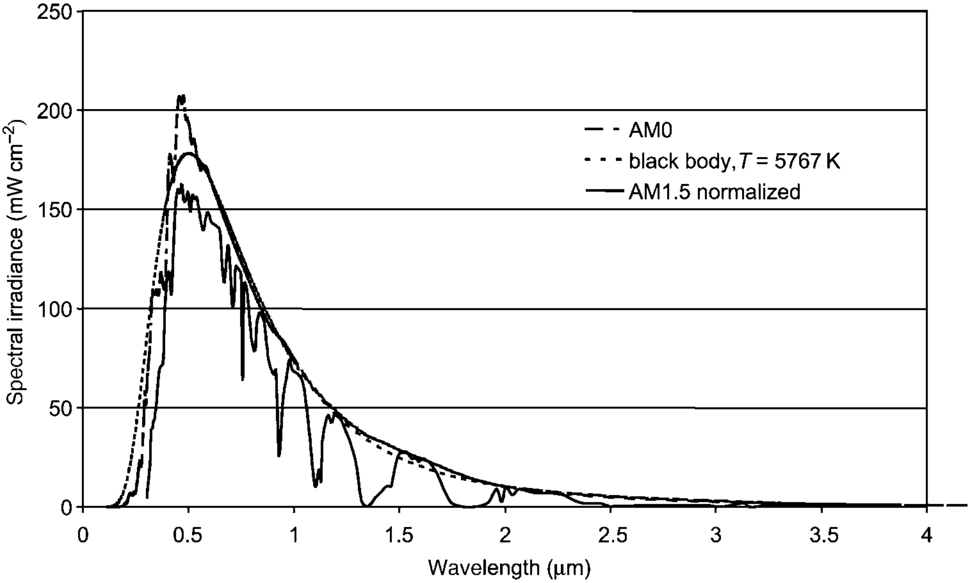


Figure C4.4.4. The different solar radiation spectra that are used in the solar cell theory and characterization.

$$J_{\mathrm{SC}} = q\int \delta\phi = \frac{2\pi}{c^2} q f_\omega \int_{\nu_g}^{\infty} \frac{\nu^2 \delta\nu}{e^{\frac{h\nu}{kT_S}} - 1}. \tag{C4.4.6}$$

A similar argument can be used to obtain the dark saturation current density J_0, but now the black body radiation (which completely surrounds the cell) is at the ambient temperature T:

$$J_0 = \frac{4\pi}{c^2} q \int_{\nu_g}^{\infty} \frac{\nu^2 \delta\nu}{e^{\frac{h\nu}{kT}} - 1}. \tag{C4.4.7}$$

Note that there is an extra factor of 2 on account of the fact that the black body radiation at ambient temperature is emitted from twice the cell area (i.e. from the front and back of the cell). The open circuit voltage can now be obtained from equations (C4.4.6) and (C4.4.7) with the aid of equation (C4.4.2). The maximum solar cell efficiency which is thus obtained depends only on the bandgap of the semiconductor from which the solar cell is made (figure C4.4.5).

Higher efficiencies can be obtained if solar radiation is 'concentrated' by means of lenses or mirrors, or if several semiconductors with different bandgaps are used to convert different parts of the spectrum. The former situation is shown in figure C4.4.5 by assuming that the cell is completely surrounded by the source of the black body radiation (i.e. by setting $f_\omega = 1$). The combination of different semiconductors to make 'tandem cells' is discussed below.

Other methods to calculate the solar efficiency for a single-gap cell include arguments based on thermodynamic principles which underline the fact that, near the open circuit, the solar cell behaves as an ideal thermodynamic engine with Carnot efficiency [14]. Using a simplified two-level model,

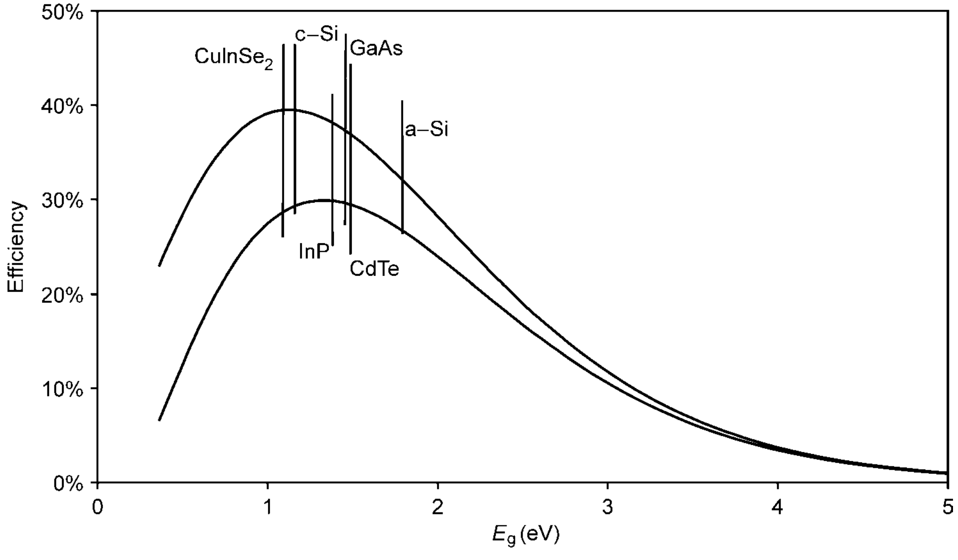


Figure C4.4.5. The Shockley–Queisser detailed balance limit on the solar cell efficiency. The lower efficiency corresponds to ordinary sunlight with geometrical factor f_ω given by equation (C4.4.4). The higher value corresponds to the case where the radiation source completely surrounds the solar cell obtained by setting $f_\omega = 1$. Also shown are the bandgaps of the most common photovoltaic materials (c-Si and a-Si denotes crystalline and amorphous silicon, respectively).

Baruch *et al* [8] showed that the maximum open circuit voltage produced by a solar cell made from semiconductor with bandgap E_g can be approximated by the expression

$$V_{oc} = \frac{E_g}{q}\left(1 - \frac{T}{T_S}\right) + \frac{kT}{q}\ln\left(\frac{f_\omega}{2}\right). \tag{C4.4.8}$$

The first term in equation (C4.4.8) is the bandgap energy in eV multiplied by the Carnot efficiency, and the second term gives the reduction in V_{oc} on account of the dilution of solar black body radiation on earth. A factor of two has been added in the denominator of the second term, in keeping with a similar term in equation (C4.4.7). A closer agreement between the methods of Baruch *et al* [8] and Shockley and Queisser [9] can be obtained by considering the solar cell as a parallel connection of two-level systems which comply with the open circuit limit (equation (C4.4.8)). Figures C4.4.6 and C4.4.7 compare the theoretical values of the short circuit current and open circuit voltage with data for the best solar cells to date from different materials.

The ideal solar cell efficiencies discussed above refer to single-junction semiconductor devices whose principal efficiency limitations are due to the inability of the semiconductor to absorb below bandgap photons, and to the fact that a part of the energy of above-bandgap photons is lost as heat. To overcome these limitations, one can form a tandem cell by stacking two or more cells on top of each other, each converting its own part of the spectrum. Devices of this form which are now available commercially include high-efficiency solar cells for satellites from Spectrolab or Tecstar in the USA, for example, or thin film amorphous silicon or silicon/germanium double or triple junctions tandem cells (for example, from Unisolar or BP Solar). The tandem cell technology may, in principle, increase the achievable efficiencies to 42% (55% under concentrated sunlight) for a double-tandem structure, or to 86.8% in the limit of an infinite number of cells [6, 15].

Further improvements are possible by a variety of means which have recently been referred to as the ‘third generation’ technologies [16]. These ideas involve, for example, the creation of more than one electron–hole pair from the incident photons—in other words, quantum efficiencies in excess of unity. This use of ‘impact ionization’ to improve the cell efficiency was first proposed by Landsberg *et al* [17].

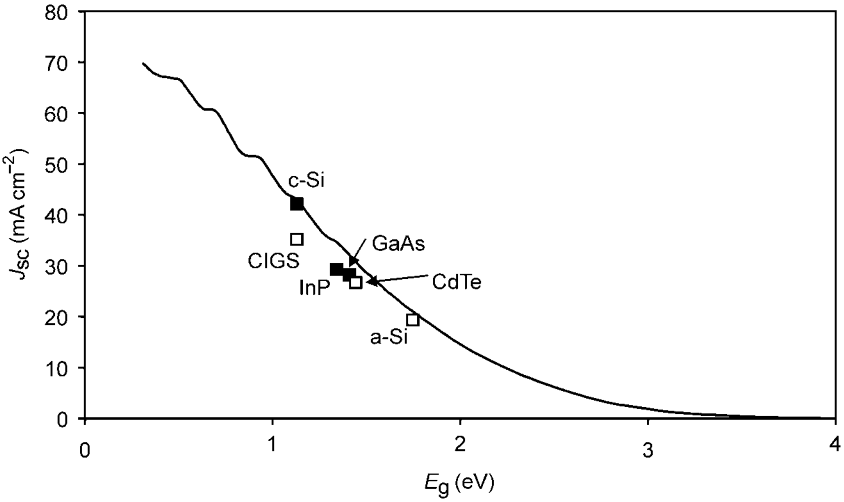


Figure C4.4.6. The maximum theoretical short circuit current density under the standard AM1.5 illumination (full line) and the best measured values for solar cells from different materials (CIGS stands for copper indium–gallium di-selenide. For other symbols, see figure C4.4.5.).

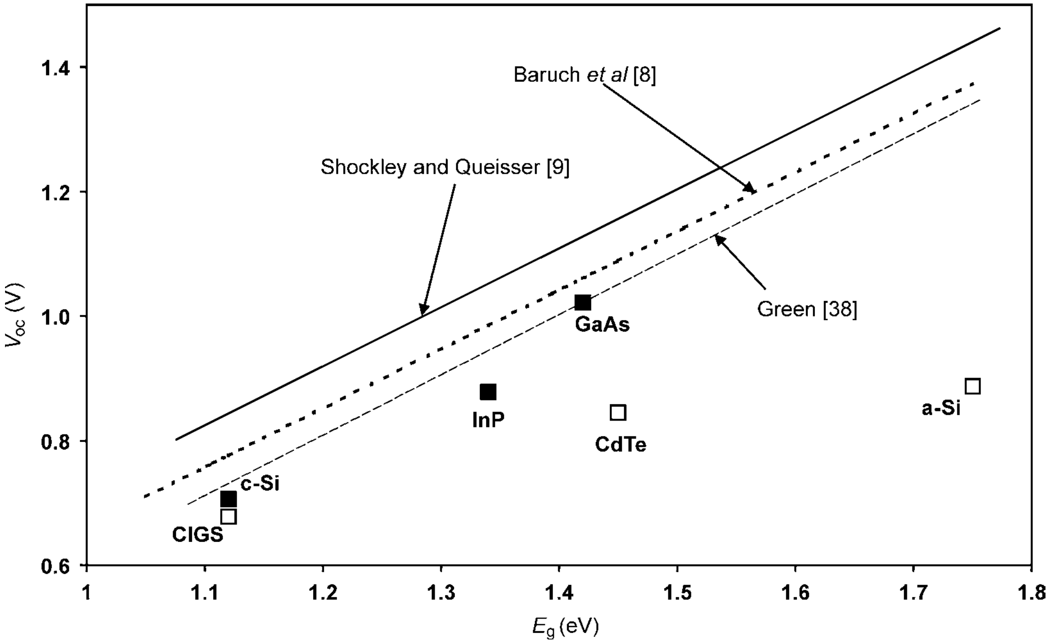


Figure C4.4.7. The best measured values of open circuit voltage for crystalline materials and thin films (■ and □, respectively) compared with theoretical predictions. The theories of Shockley and Queisser and Baruch *et al* are discussed in the text. The theory of Green [38] gives a semi-empirical estimate of the maximum achievable voltage.

Kolodinski *et al* [18] demonstrated the feasibility of this approach but a working solar cell is yet to be developed. Other theoretical third-generation concepts invoke hot electrons [19, 20], or the use of impurities [21, 22] or intermediate bands [23, 24] to utilize the below-bandgap light. Quantum wells have also been suggested as a means of improving efficiency [25]. (Nelson [26] gives a recent review of this field.) A thermodynamic efficiency limit for this general form of solar energy converter is known as the Landsberg efficiency [27, 28].

C4.4.3 Solar cell technologies

The photovoltaic market today is dominated by crystalline silicon, in the single crystal or multicrystalline form (figure C4.4.8). Amorphous silicon makes up most of the remaining production, whilst other technologies, including the thin films CdTe and Cu(In,Ga)Se_2, and solar cells based on the ribbon or nonwafer silicon, are at the start of commercial production. There is also a small market for solar cells for concentrator systems which can make use of expensive high-efficiency solar cells whose application has so far remained in the domain of the space technology.

Space solar cells are not included in figure C4.4.8 although their production—much smaller now in energy terms—is not insignificant in terms of financial turnover. Although similar in structure, the main requirements on solar cells used to power satellites in space differ substantially from their terrestrial counterparts. The primary drivers for space cells are the power output (or more precisely, the power-to-weight ratio) and, in many instances, a good radiation resistance. The latter requirement is, in fact, the main reason for the predominant n-on-p configuration of today's silicon solar cells, since the p-type base where most of the power is generated, is more radiation resistant than the n-type material.

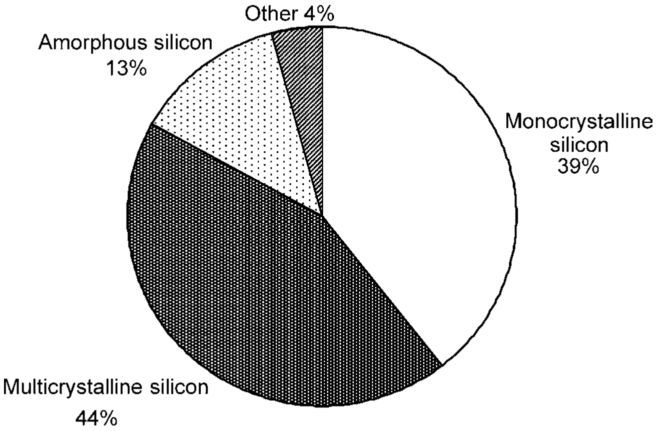


Figure C4.4.8. The market share of different photovoltaic technologies. 'Other technologies' include crystalline silicon cells for concentrator systems, cells based on ribbon silicon, cadmium telluride and silicon on ceramics. (Courtesy P Maycock, PV News. Data for 1998.)

The principal effect of particle radiation on the solar cell comes through the change in minority-carrier lifetime and the resultant reduction in the diffusion length. Over the recent years, the requirements for high efficiency and radiation tolerance have swayed the balance in favour of GaAs, sometimes in the form of a double junction structures based on GaInP/GaAs. Indium phosphide cells have been suggested for use in harsh radiation environments as their radiation resistance is particularly high, due to partial annealing of this damage under illumination [29].

C4.4.3.1 Crystalline silicon

A typical monocrystalline silicon solar cells in production today has efficiency around 15%, slightly less for cells made from the multicrystalline material. It is typically made by phosphorus diffusion of the top layer (emitter) into a p-type wafer, with screen printed metal contacts and a thin layer of anti-reflection coating (arc; figure C4.4.9) [30]. Photons, absorbed mainly in the base of the cell, create electron–hole pairs. These are separated by the electric field of the junction (figure C4.4.10(*a*)). The principal efficiency losses incurred in such a cell are shown in figure C4.4.11.

Improvements to the basic structure can be made by texturing the top surface to reduce the optical reflection. In combination with an optically reflecting rear surface, this can be used to produce a significant degree of light trapping to off-set the poor optical absorption of silicon. Surfaces can be passivated to reduce surface recombination, and p^+ diffusion can be used to create a back-surface field (a barrier to minority carrier transport) and reduce recombination near the back surface. Laser-grooved buried-contact technology, invented by Martin Green at the University of New South Wales in Australia and utilized by the BP Solar 'Saturn' cells, reduces the shading by top surface metalization and pushes up the efficiency of production cells close to 18%. In the laboratory, the top silicon solar cell efficiency has reached almost 25%—about 80% of the theoretical maximum for single junction devices. Two high-efficiency structures which have achieved record conversion efficiencies are shown in figure C4.4.12. The Stanford cell, with both sets of contacts at rear to avoid shading, is intended for operation under concentrated sunlight.

Silicon is an indirect-gap material, and the low absorption coefficient necessitates relatively thick devices to make good use of the available sunlight (figure C4.4.13). This is most easily achieved by

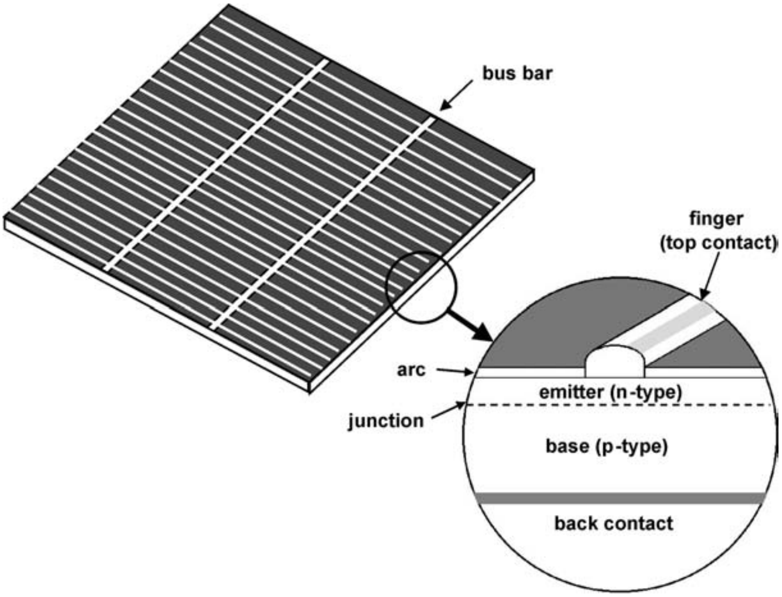

bus bar
finger
(top contact)
arc
junction
emitter (n-type)
base (p-type)
back contact

Figure C4.4.9. The structure of a typical crystalline silicon solar cell in production today.

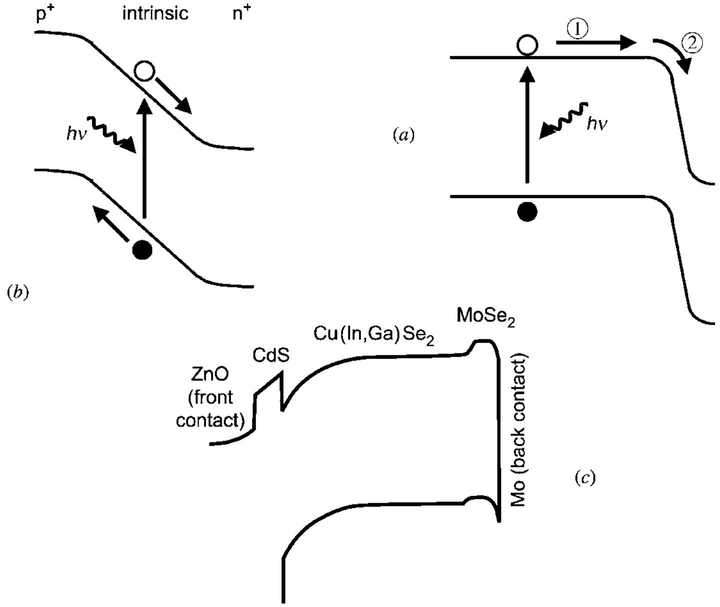

p+
intrinsic
n+
hν
(b)
(a)
①
②
hν
ZnO
(front
contact)
CdS
Cu(In,Ga)Se2
MoSe2
Mo (back contact)
(c)

Figure C4.4.10. (*a*) A schematic diagram of the band structure and operation of a crystalline silicon solar cell. Most minority carriers are generated in the p-type base, and must diffuse to the junction (arrow 1) before charge separation (arrow 2). (*b*) In a typical amorphous silicon p–i–n structure, electron–hole pairs are generated in the space charge region where the built-in field aids carrier collection. (*c*) The band structure of a $Cu(InGa)Se_2$ solar cell (after [33]).

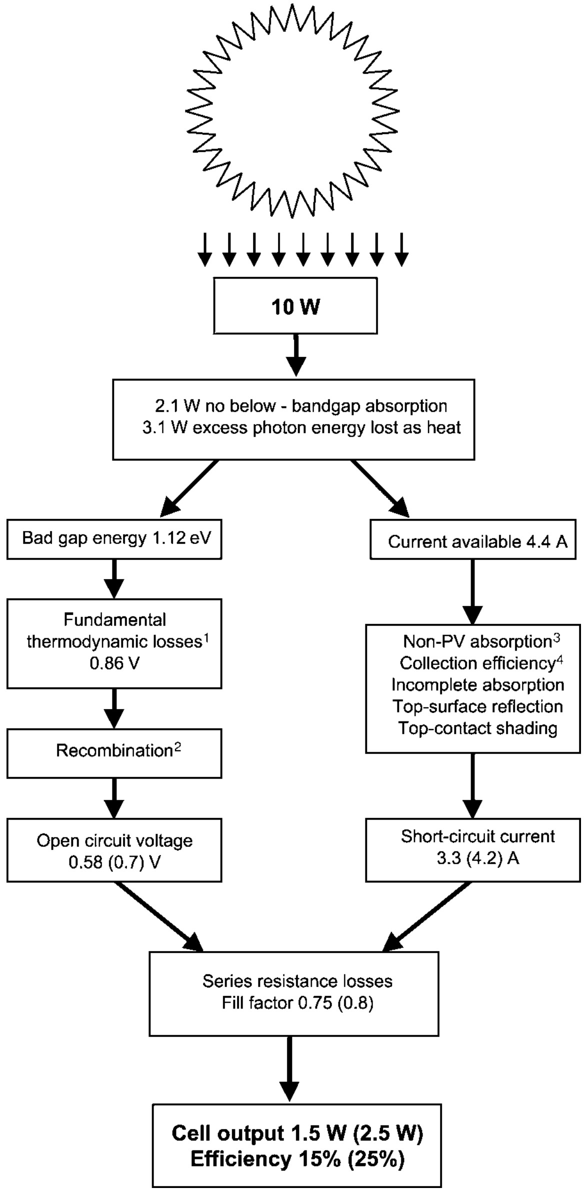
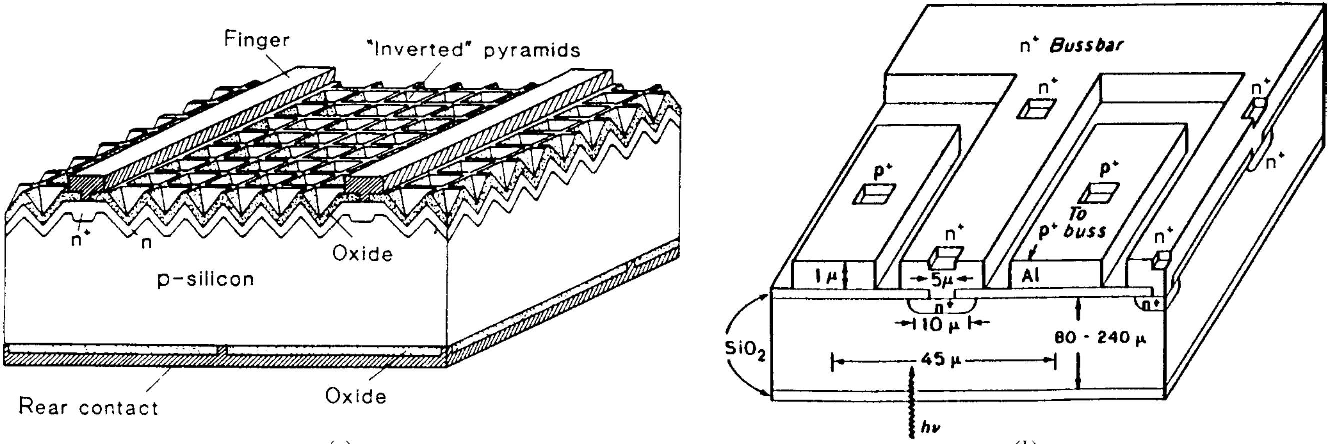

10 W
2.1 W no below - bandgap absorption
3.1 W excess photon energy lost as heat
Bad gap energy 1.12 eV
Current available 4.4 A
Fundamental thermodynamic losses1 0.86 V
Non-PV absorption3
Collection efficiency4
Incomplete absorption
Top-surface reflection
Top-contact shading
Recombination2
Open circuit voltage 0.58 (0.7) V
Short-circuit current 3.3 (4.2) A
Series resistance losses
Fill factor 0.75 (0.8)
Cell output 1.5 W (2.5 W)
Efficiency 15% (25%)

Figure C4.4.11. Principal efficiency losses in a commercial 100 × 100 mm crystalline silicon solar cell, compared with those for the PERL cell (in brackets, see figure C4.4.12). Notes: 1. See equation (C4.4.8). Losses due to dilution of solar radiation (the second term) are absent under maximum concentration of sunlight. 2. Surface and volume recombination in the base, emitter and junction regions. 3. Principally free-carrier absorption. 4. Collection efficiency is limited mainly by recombination in the base and emitter (after [43]).


Finger
"Inverted" pyramids
n+
n
Oxide
p-silicon
Rear contact
Oxide
n+ Bussbar
n+
n+
n+
p+
p+
n+
p+
To buss
n+
n+
1 μ
5μ
Al
n+
n+
10 μ
80 - 240 μ
SiO2
45 μ
hν

(*a*) (*b*)

Figure C4.4.12. The UNSW PERL cell [39] and the Stanford point-contact cell [40].

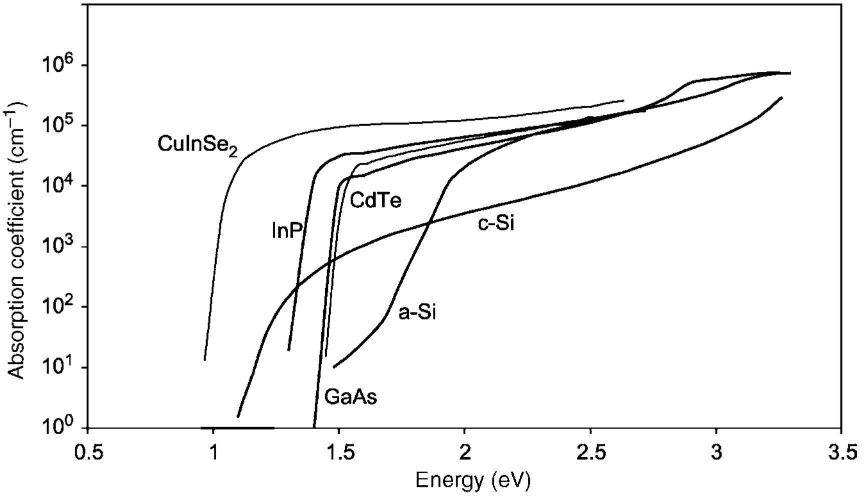


Figure C4.4.13. The absorption coefficients of selected photovoltaic materials.

a wafer-based technology. The major part of crystalline silicon cell production is made from wafers cut from monocrystalline or multicrystalline ingots. To avoid the kerf loss, a number of companies (including ASE, Evergreen Solar and Ebarra) are turning to noningot wafer technologies such as dendritic or edge-defined film growth (EFG) [31]. Other fabrication methods which use less silicon include the growth of thin crystalline silicon solar cells deposited on a ceramic substrate, as demonstrated by Astropower. Since the early days of photovoltaics, however, there has been a substantial effort directed towards finding a cheap replacement material for crystalline silicon—for a semiconductor with good optical properties which can be deposited as a thin-film to reduce the materials requirements.

C4.4.3.2 Thin film solar cells

The origin of thin film solar cells dates back to the same year as crystalline silicon solar cells [32]. Commercial production of the first thin film devices had to await the development of hydrogenated amorphous silicon (a-Si) in the second half of the 1970s. The discovery of a-Si solar cell degradation associated with the Staebler–Wronski effect [45], however, represented a major setback to the new technology, and most a-Si solar cells produced today implement some measures to reduce its impact.

The present single junction a-Si cells (used principally to power commercial products such as watches and calculators) are invariably based on p–i–n or n–i–p technology. The incorporation of the intrinsic layer serves to enhance carrier collection since most electron–hole pairs are generated in a region where electric field assists their separation (figure C4.4.10). The manufacture involves first the deposition of a layer of tin oxide on glass to act as a transparent front contact. The p–i–n cell is then formed by depositing a p-layer of boron doped silicon (or silicon/silicon carbide alloy), intrinsic silicon, the phosphorus-doped n-layer, and finally a metal back contact. Tandem a-Si cells which are increasingly more common sometimes incorporate also a-Si/a-Ge layers to optimize the bandgap. Other novel structures involve the combination of microcrystalline/amorphous silicon solar cells. Two versions of such a device are marketed by Kaneka Corporation and Sanyo of Japan.

Other thin film solar cell technologies based on compound semiconductors—cadmium teluride and copper indium diselenide or its derivatives—are currently aiming to start or scale up commercial

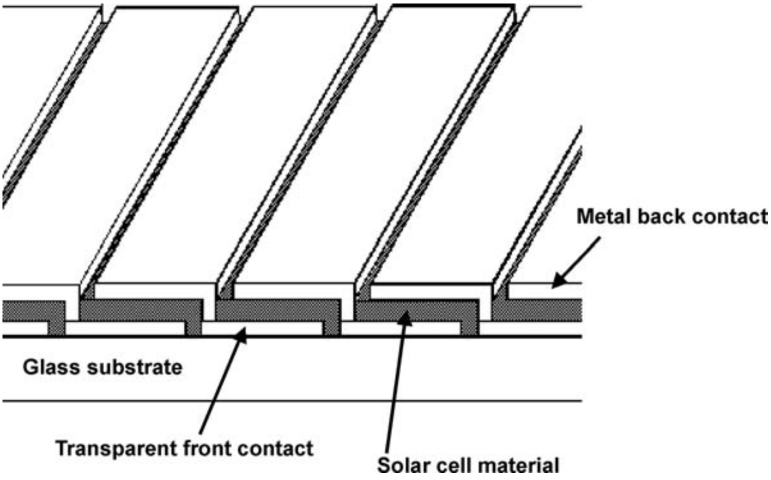


Figure C4.4.14. A schematic diagram of an integrally interconnected module [41].

production. The CdTe cells are usually manufactured in the form of CdTe/CdS heterojunctions. Matsushita Corporation of Japan has been producing CdTe cells for some time, joined more recently by BP Solar manufacture in California. First Solar LLC of Toledo, Ohio is scaling up pilot line production and ANTEC Solar is about to start production in Germany.

The other main compound semiconductor which is being used in commercial production of solar cells is copper indium diselenide. These cells are often denoted by CIGS as they now usually employ $Cu(In,Ga)Se_2$ to optimize the bandgap, and have achieved the highest research efficiencies among the thin film technologies, approaching 19% [33]. The band structure of a typical CIGS solar cell, based on a $CdS/CuInGaSe_2$ heterostructure, is shown in figure C4.4.10(*c*). Siemens Solar, Global Solar, EPV and Würth Solar (using technology developed at Stuttgart's Solar Hydrogen Centre, ZSW) have recently started commercial production of CIS PV modules.

An advantage of thin-film solar cells over the wafer-based technologies lies not only in the lower material requirements but also in the possibility of integrated module manufacture. Alternating layers can be deposited and laser scribed with an offset in such a way that a series connection of cells is produced without mechanical handling of individual cells (figure C4.4.14). This helps to reduce the cost still further but the relatively lower efficiencies of thin film cells, combined often with stability problems, have so far hindered their use in large scale installations.

C4.4.3.3 Dye-sensitized and organic solar cells

The search for new materials includes the development of devices where the photovoltaic energy conversion is carried out by molecules instead of semiconductor structures. In the early 1990s, Michael Grätzel and colleagues at the Ecole Polytechnique Fédérale de Lausanne demonstrated a solar cell where Ru-based molecular dyes both absorb the light and participate in charge separation (figure C4.4.15). In effect, the mono molecular dye layer 'pumps' electrons from one electrode (liquid electrolyte) to the other (solid titanium dioxide). The use of nanocrystalline titanium oxide particles coated by the dye ensures high optical absorption: virtually all light within the spectral range of the dye is absorbed by a coating not more than few angstroms thin.

There is also considerable research activity into true molecular structures. A number of research groups have demonstrated solar cells based on polymers, and there is much to look forward to if their

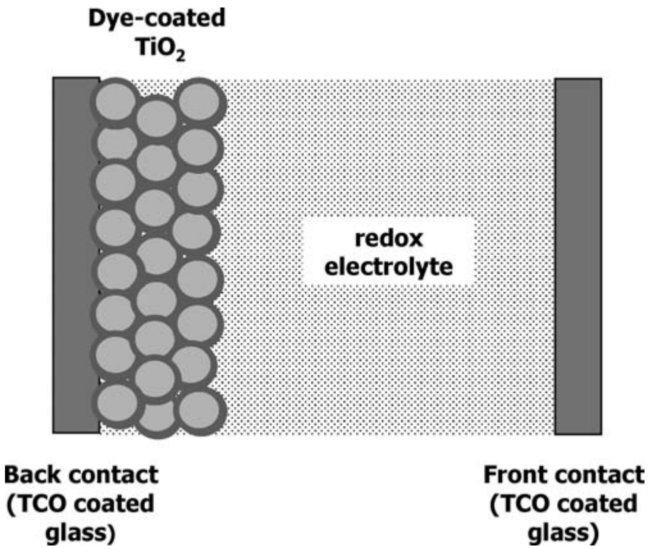

Dye-coated TiO₂
redox electrolyte
Back contact (TCO coated glass)
Front contact (TCO coated glass)

Figure C4.4.15. A schematic diagram of the Grätzel dye-sensitized solar cell (after [42]).

success matches the achievements of the LED technology. These developments often make use of a combination of conducting conjugated polymers and fulerene derivatives as electron acceptors to create a p–n junction.

C4.4.4 Issues in system design

For practical use, solar cells are laminated and encapsulated to form photovoltaic modules. These are then combined into arrays, and interconnected with other electrical and electronic components—for example, batteries, charge regulators and inverters—to form a photovoltaic system. A number of issues need to be resolved before an optimum system design is achieved. These issues include the choice between a flat plate or concentrating system, and the required array configuration (fixed tilt or tracking). Answers to these questions will vary depending on the solar radiation at the site of the installation. Specific issues also relate to whether the system is to be connected to the utility supply (the 'grid') or is intended for stand-alone operation.

The available solar radiation, of course, varies from day to day, season to season, and depends on the geographical location. One can, however, obtain an estimate for the mean daily irradiation (solar energy received by a unit area in 1 day), averaged over the surface of the globe. To this end, we note that the total energy flux incident on the earth is equal to the solar constant S multiplied by the area of the disc presented to the sun's radiation by the earth. The average flux incident on a unit surface area is then obtained by dividing this number by the total surface area of the earth. Making allowance for 30% of the incident radiation being scattered and reflected into space, the average daily solar radiation G_d on the ground is equal to

$$G_\mathrm{d} = 24 \times 0.7 \times \frac{\pi R_\mathrm{E}^2}{4\pi R_\mathrm{E}^2} \times S = 24 \times 0.7 \times \frac{1}{4} S = 5.74\,\mathrm{kWh} \qquad \text{(C4.4.9)}$$

where R_E is the radius of the earth which was introduced in section C4.4.2. The value should be compared with the observed values. Figure C4.4.16 shows the daily solar radiation on a horizontal plane for four

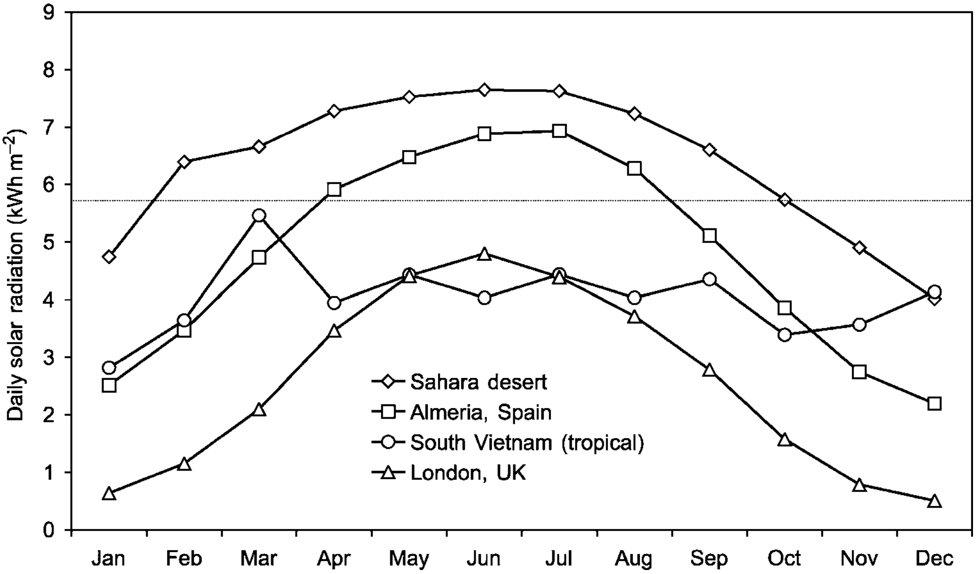


Figure C4.4.16. The mean daily solar radiation and its monthly variation in different regions of the world. The dashed line indicates the average value (equation (C4.4.9)).

locations ranging from the humid equatorial regions to northern Europe. The solar radiation is highest in the continental desert areas around latitude 25–30° N and 25–30° S, and falls off towards the equator because of the cloud, and towards the poles because of low solar elevation. Equatorial regions experience little seasonal variation, in contrast with higher latitudes where the summer/winter ratios are large.

Most photovoltaic arrays are installed at fixed tilt and, wherever possible, oriented towards the equator. The optimum tilt angle is usually determined by the nature of the application. Arrays which are to provide maximum generation over the year (for example, many grid-connected systems) should be inclined at an angle equal to the latitude of the site. Stand-alone systems which are to operate during the winter months have arrays inclined at a steeper angle of latitude +15°. If power is required mainly in summer (for example, for water pumping and irrigation), the guide inclination is latitude −15°.

The amount of solar energy captured can be increased if the modules track the sun. Full two-axis tracking, for example, will increase the annual energy available by over 30% over a nontracking array fixed at the angle of latitude (figure C4.4.17)—at the expense, however, of increased complexity. Single axis tracking is simpler but yields a smaller gain. Tracking is particularly important in systems which use concentrated sunlight. These systems can partially offset the high cost of solar cells by the use of inexpensive optical elements (mirrors or lenses). The cells, however, then usually need to be cooled and it should also be borne in mind that only direct (beam) solar radiation can be concentrated to a significant degree, thus reducing the available energy input. This effectively restricts the application of concentrator systems to regions with clear skies.

There is a considerable difference between the design of stand alone and grid connected systems. Much of the difference stems from the fact that the design of stand-alone systems endeavours to make the most of the available solar radiation. This consideration is less important when utility supply is available, but the grid connection imposes its own particular constraints which must be allowed for in the system design.

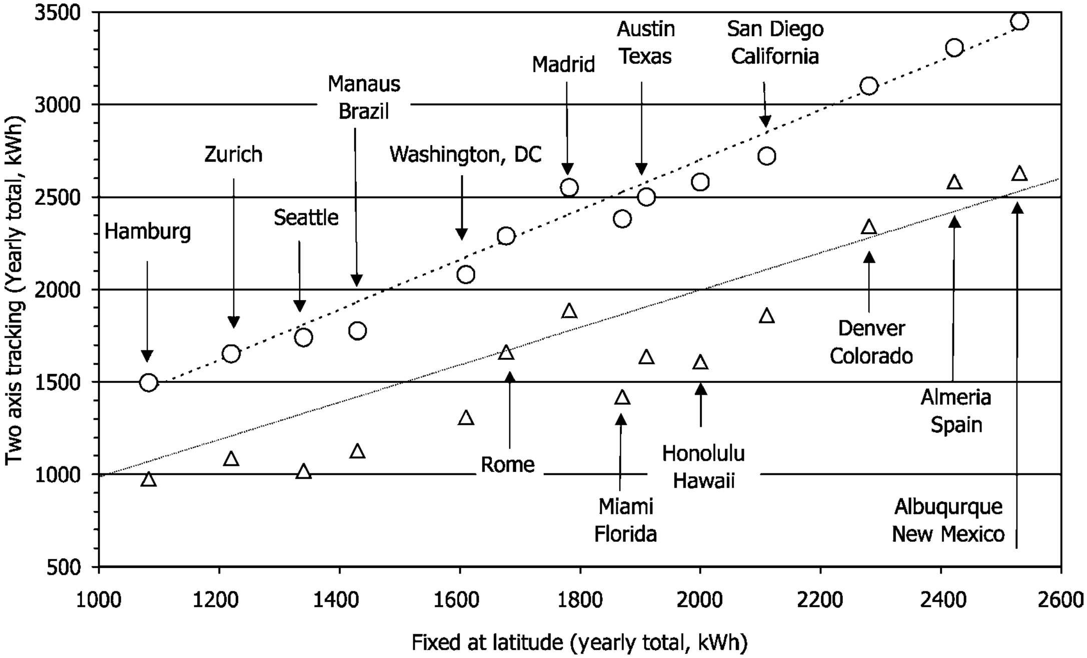


Figure C4.4.17. The energy available to a photovoltaic array with two-axis tracking, in comparison with a flat plate system inclined at an angle equal to the latitude of the site. Global radiation is shown by ○, direct beam by Δ. The full and dotted lines correspond to the global radiation available to the fixed array and a 34% increase over this value, respectively (after [43]).

C4.4.4.1 Stand-alone systems

The various applications of PV systems in isolated locations have already been mentioned in section C4.4.1. The diversity of uses leads to different requirements on the system and different system specifications. An important parameter that reflects the nature of the application is the required security of supply. Telecommunication and systems used for marine signals, for example, need to operate at a very high level of reliability. In other applications, the user may be able to tolerate lower reliability in return for a lower system cost [34]. These considerations have an important bearing on how large PV array and how large energy storage (battery) need to be installed, in other words, on system sizing.

Among the variety of sizing techniques, sizing based on energy balance provides a simple and popular technique which is often used in practice. It gives a simple estimate of the PV array necessary to supply a required load, based on an average daily solar irradiation at the site of the installation, available now for many locations in the world. Choosing the month with the lowest irradiation (for example, December in northerly latitudes) and the value appropriate to the inclined panel, the energy balance equation can then be found with the use of data in figure C4.4.16:

$$\text{Array size (in W}_\text{p}) = \frac{\text{Daily energy consumption}}{\text{Daily solar radiation}}. \qquad \text{(C4.4.10)}$$

Equation (C4.4.10) specifies how many PV modules need to be installed to supply the load under average conditions of solar irradiation. The battery size is then estimated ‘from experience’—the rule of thumb

Table C4.4.2. Area and nominal power of a photovoltaic array, facing south and inclined at latitude angle, needed to produce 1 MWh of electricity per year at different locations, for two module efficiencies η.

Location	Nominal array power (kW_p)	Module area (m^2)	
		$\eta = 15\%$	$\eta = 10\%$
Sahara desert	0.41	2.7	4.1
Almeria, Spain	0.53	3.5	5.3
London	1.04	6.9	10.4
Vietnam	0.65	4.4	6.5

recommends, for example, installing 3 days of storage in tropical locations, 5 days in southern Europe and 10 days or more in the UK. As an illustration, table C4.4.2 shows the size of the PV array needed to generate 1 MWh per year in different locations around the world.

Although easy to use, the sizing method based on energy balance does not give an indication how the PV system will operate under fluctuating solar radiation. In particular, it does not predict the reliability of energy supply. The relationship between the reliability of supply and sizing is illustrated well by the random walk method [35]. The method consists of treating the possible states of charge of the battery as discrete numbers which are then identified as sites for a random walk. Each day, the system makes a step in the ransom walk depending on solar radiation: one step up if it is ‘sunny’ and one step down if it is ‘cloudy’. The magnitudes of these steps and probabilities of weather being ‘sunny’ and ‘cloudy’ are determined from the solar radiation data and the daily load. When the random walker resides in the top state, the battery is fully charged; when it is in the bottom state, the battery is completely discharged and the load is disconnected. The calculations are carried out by assuming that, after a certain time, the random walker reaches a steady state. The loss of load probability (LLP) is then equated to the probability of the random walker residing in the lowest state. Bucciarelli [36] subsequently extended this method to allow for correlation between solar radiation on different days. For a given LLP, the result can be expressed in the form of a sizing curve: a functional dependence of the array size on the battery size from which the least-cost system can easily be determined (figure C4.4.18). These and other more complex sizing techniques have been summarized in more detail by Gordon [37].

C4.4.4.2 Grid-connected systems

Grid-connected systems have grown considerably in number since the early 1990s spurred by government support programmes for ‘photovoltaics in buildings’ in a number of countries, led initially by Switzerland and followed by more substantial programmes in Germany and Japan. One feature that affects the system design is the need for compliance with the relevant technical guidelines to ensure that the grid connection is safe; the exported power must also be of sufficient quality and without adverse effects on other users of the network. Although a common set of international standards is still some way off, it is probably fair to say that the fundamental issues have been identified—partly through the work of the Task 5 of the International Energy Agency. In a number of countries, the required statutory guidelines have now been produced: in the UK, for example, the relevant Engineering Recommendation G77 was published in 2000. An example of the requirements imposed on the grid interface of a photovoltaic system by the utility is shown in table C4.4.3.

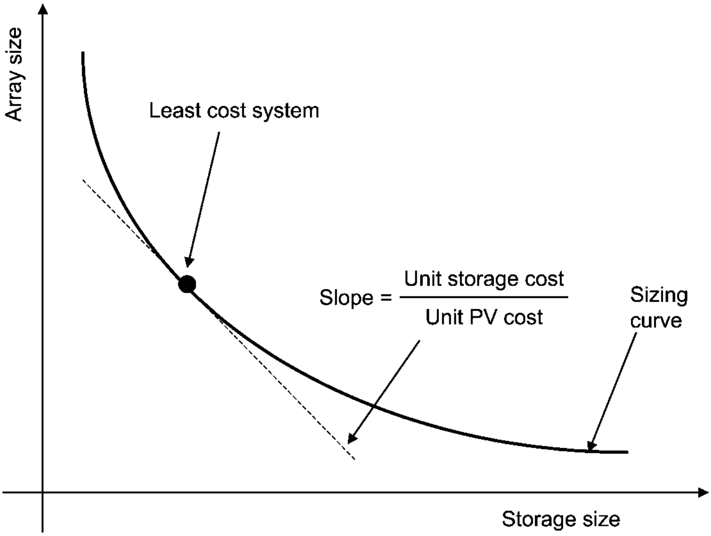


Figure C4.4.18. The sizing curve, representing the locus of PV system configurations with the same reliability of energy supply. The least-cost system can be found by a simple geometrical construction.

Many of the grid connection issues are not unique to photovoltaics. They arise from the difficulties trying to accommodate 'embedded' or distributed generators in an electricity supply system designed around large central power stations. It is likely that many of these features of grid connection will undergo a review as electricity distribution networks evolve to absorb a higher proportion of embedded

Table C4.4.3. A summary of principal requirements on the grid interface of PV generators, covered in the UK Engineering Recommendation G77.

Function	Reference
Protection	
General	IEC 255
Under/over voltage	UK guidelines[a]
Under/over frequency	UK guidelines[a]
Loss of mains	Specific for PV
Supply quality	
Harmonics	BS and EN[b]
Voltage flicker	BS and EN[b]
Electromagnetic compatibility	BS and EN[b]
DC injection	UK guidelines[a]
Safety	
Earthing	BS[c]

[a] Covered by existing UK guidelines for parallel connection of embedded generators.
[b] Covered by existing British standard and European norm.
[c] Covered by existing British standard.

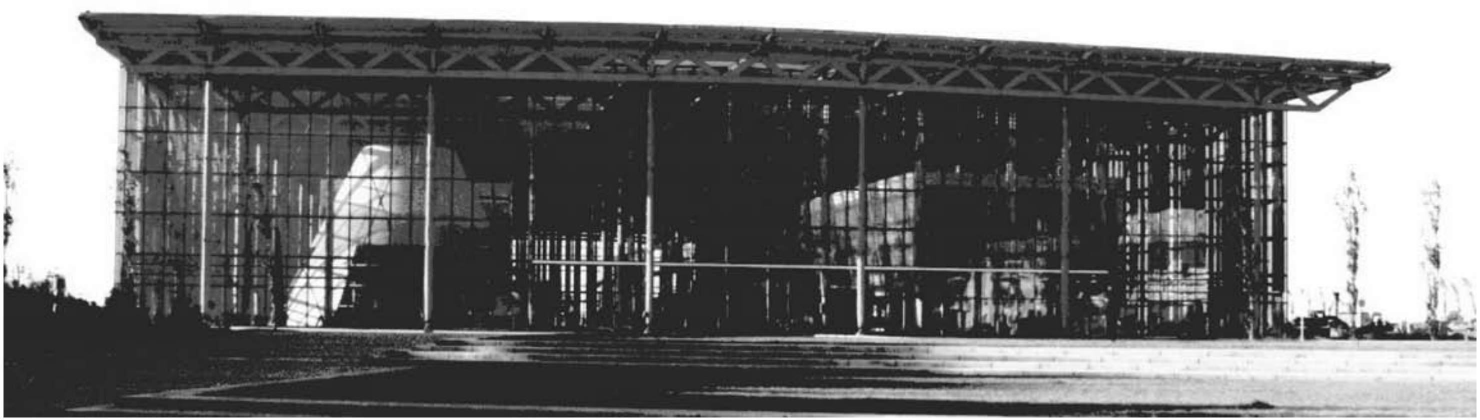


Figure C4.4.19. The Mont Cenis Academy at Herne-Sodingen.

generation: wind farms, co-generation (or CHP) units, or other local energy sources. The electricity supply system in 20 or 30 years time might be quite different from now, and new and innovative integration schemes will be needed to ensure optimum integration. Photovoltaic generators are likely to benefit from these changes, particularly from the recent advances in the technology of small domestic size co-generation units (micro-CHP) which have a good seasonal synergy with the energy supply from solar sources, and can share the cost of the grid interface.

An example of how elegant architecture can be combined with forward looking engineering is offered by the Mont-Cenis Energy Academy at Herne-Sodingen (figure C4.4.19). This solar-cell clad glass envelope at the site of a former coal mine provides a controlled Mediterranean microclimate which is powered partly by 1 MW_p photovoltaic array and partly by two co-generation units fuelled by methane released from the disused mine. To ensure a good integration into the local electricity supply, the generators are complemented by a 1.2 MWh battery bank. In addition to the Academy, the scheme also exports electricity and heat to 250 units in a nearby housing estate and a local hospital. The Mont-Cenis Academy is a fine flag-carrier for photovoltaics and new energy engineering—without a doubt, similar schemes will become more prolific as photovoltaics and energy efficient solutions become the accepted norm over the next few decades.

C4.4.5 Conclusions

Photovoltaic technology has come a long way since the first solar powered satellites in the late 1950s. Based today mainly on crystalline silicon, new approaches which utilize thin-film technologies and molecular materials are beginning to make their mark in commercial production. If the current rate of growth continues, distributed solar power systems integrated into buildings will supply a significant part of our energy needs during the early part of this new millennium. One can foresee a bright future for this new, clean, energy source.

Acknowledgments

The author is grateful to Dr Neil Ross for his careful reading of the manuscript and a number of useful comments.

References

[1] Becquerel A E 1839 Recherches sur les effets de la radiation chimique de la lumière solair au moyen des courants électriques *C. R. Acad. Sci.* **9** 561–567

[2] Smith W 1873 The action of light on selenium *J. Soc. Telegraph Eng.* **2** 31–33

[3] Adams W G and Day R E 1877 The action of light on selenium *Proc. R. Soc.* **A25** 113–117

[4] Chapin D M, Fuller C S and Pearson G O 1954 A new silicon p–n junction photocell for converting solar radiation into electrical power *J. Appl. Phys.* **25** 676–677

[5] Landsberg P T and Badescu V 1998 Solar Energy conversion: list of efficiencies and some theoretical considerations *Prog. in Quantum Electron.* **22** 211–230

[6] deVos A 1992 *Endoreversible Thermodynamics of Solar Energy Conversion* (Oxford: Oxford University Press)

[7] Würfel P 1982 The chemical potential of radiation *J. Phys.* **C15** 3967–3985

[8] Baruch P, Picard C and Swanson R M 1980 Thermodynamical limits to photovoltaic solar energy conversion efficiency *Proc. 3rd European Photovoltaic Energy Conf.* p 927

[9] Shockley W and Queisser H J 1961 Detailed balance limit of efficiency of pn junction solar cells *J. Appl. Phys.* **32** 510–519

[10] Müser H A 1957 Thermodynamische Behandlung von Elektronenprozessen in Halbleiter-Randschiten *Z. Physik* **148** 380–390

[11] Hulstrom R, Bird R and Riordan C 1985 Spectral solar irradiance data sets for selected terrestrial conditions *Solar Cells* **15** 365–391

[12] Thekaekara M P 1977 *Solar Energy Engineering*, ed A A M Sayigh (New York: Academic) p 37

[13] Parrott J E 1993 Choice of an equivalent black body solar temperature *Solar Energy* **51** 195

[14] Baruch P and Parrott J E 1990 A thermodynamic cycle for photovoltaic energy conversion *J. Phys. D: Appl. Phys.* **23** 739

[15] deVos A 1980 Detailed balance limit of the efficiency of tandem solar cells *J. Phys. D: Appl. Phys.* **13** 839

[16] Green M A 2001 Third generation photovoltaics: Ultra high conversion efficiency at low cost *Prog. Photovoltaics Res. Appl.* **9** 123–135

[17] Landsberg P T, Nussbaumer H and Willeke G 1993 Band-band impact ionization and solar energy efficiency *J. Appl. Phys.* **74** 1451–1452

[18] Kolodinski S, Werner J H, Wittchen T and Queisser H J 1993 Quantum efficiencies exceeding unity due to impact ionisation in silicon solar cells *Appl. Phys.* **63** 2405

[19] Ross R T and Nozik A J 1982 Hot-carrier solar energy converters *J. Appl. Phys.* **53** 3813–3818

[20] Würfel P 1997 Solar energy conversion with hot electrons from impact ionisation *Solar Energy Mater. Solar Cells* **46** 43

[21] Würfel P 1993 Limiting efficiency for solar cells with defects from three-level model *Solar Energy Mater. Solar Cells* **29** 403–413

[22] Keevers M and Green M A 1994 Efficiency improvements of silicon solar cells by the impurity photovoltaic effect *J. Appl. Phys.* **75** 4022–4031

[23] Luque A and Martí A 1997 Increasing the efficiency of ideal solar cells by photon induced transitions at intermediate levels *Phys. Rev. Lett.* **78** 5014–5017

[24] Luque A and Martí A 2001 A metallic intermediate band high efficiency solar cell *Prog. Photovoltaics Res. Appl.* **9** 73–86

[25] Barnham K W J and Duggan G 1990 A new approach to high-efficiency multi-band gap solar cells *J. Appl. Phys.* **67** 3490–3493, but see also Luque A and Martí A 1997 Entropy production in photovoltaic conversion *Phys. Rev.* **B55** 6994–6999

[26] Nelson J 2001 Quantum well solar cells *Clean Electricity from Photovoltaics*, ed M A Archer and R Hill (Singapore: World Scientific) pp 447–480

[27] Petela R 1964 Energy of heat radiation *Trans. ASME Heat transfer* **36** 187

[28] Landsberg P T and Mallinson J R 1976 Thermodynamic constraints, effective temperatures and solar cells *Int. Colloquium on Solar Electricity* (CNES, Toulouse) p 46

[29] Coutts T J and Yamaguchi M 1988 Indium phosphide based solar cells: A critical review of their fabrication, performance and operation *Curr. Top. Photovoltaics* vol 3, ed T J Coutts and J D Meakin pp 79–234

[30] Cuevas A 2000 *Crystalline silicon technology, in Solar Electricity*, ed T Markvart (Chichester: Wiley) pp 46–62

[31] Bruton T M 2002 General trends about photovoltaics based on crystalline silicon *Solar Energy Mater. Solar Cells* **72** 3–10

[32] Reynolds D C 1954 Photovoltaic effect in cadmium sulfide *Phys. Rev.* **2** 31–33

[33] Rau U and Schock H W 2001 Cu(In,Ga)Se$_2$ solar cells *Clean Electricity from Photovoltaics*, ed M A Archer and R Hill (Singapore: World Scientific) pp 277–346

[34] Lorenzo E 1997 Photovoltaic rural electrification *Prog. Photovoltaics Res. Appl.* **5** 3–27

[35] Bucciarelli L L 1984 Estimating loss-of-power probabilities of stand-alone photovoltaic solar-energy systems *Solar Energy* **32** 205–209

[36] Bucciarelli L L 1986 The effect of day-to-day correlation in solar-radiation on the probability of loss-of-power in a stand-alone photovoltaic energy system *Solar Energy* **36** 11–14

[37] Gordon J M 1987 Optimal sizing of stand-alone photovoltaic solar power systems *Solar Cells* **20** 295–313

[38] Green M A 1982 *Solar Cells* (New York: Prentice Hall)

[39] Zhao J, Wang A, Altermatt P and Green M A 1995 24% efficient silicon solar cells with double layer antireflection coatings and reduced resistance loss *Appl. Phys. Lett.* **66** 3636–3638

[40] Sinton R A, Kwark Y, Gan J Y and Swanson R M 1986 27.5-percent silicon concentrator solar cells *IEEE Electron. Device Lett.* **7** 567–569

[41] Hill R 2000 Thin film solar cells *Solar Electricity* 2nd edn, ed T Markvart (Chichester: Wiley) pp 73–74

[42] McEvoy A 2000 *Solar Electricity* 2nd edn, ed T Markvart (Chichester: Wiley) pp 247–259

[43] Boes E C and Luque A 1993 Photovoltaic concentrator technology *Renewable Energy: Sources for Fuels and Electricity*, ed T B Johansson, H Kelly, A K N Reddy and R H Williams (London: Earthscan) pp 361–401

[44] Markvart T 2000 *Solar Electricity* (Chichester: Wiley) p 43

[45] Staebler L D and Wronski C R 1980 Reversible conductivity change in discharge produced amorphous silicon *J. Appl. Phys.* **51** 3262

C4.5

Medical applications of photonics

Tuan Vo-Dinh

C4.5.1 Introduction

In the past 20 years, photonic techniques have had a dramatic effect on many different fields of research related to biomedical diagnostics and therapy [1]. Following the genomic research advances of the post-sequencing era, there has been an exploding interest in the development and applications of biochip technologies. Information on genomic sequence can be used experimentally with high-density DNA microarrays that allow complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. DNA microarrays can be used for many different purposes, especially to measure levels of gene expression (messenger RNA abundance) for tens of thousands of genes simultaneously [2–6]. On the other hand, portable, self-contained biochips with integrated detection microchip systems have great potential for use by the physician at the point of care [7–13].

Molecular spectrometry techniques—including fluorescence, scattering, absorption, reflection and optical coherence spectroscopies—have been applied to the analysis of many different types of samples, ranging from individual biochemical species (e.g. NADH, tryptophan) to organs of living animals and humans. These studies have given rise to new methods for the early or noninvasive diagnosis of various medical conditions, including tooth decay, atherosclerosis, heart arrhythmia, cancer and many others. Molecular spectrometry and/or imaging have been investigated for the diagnosis of almost every type of cancer and early neoplastic differences found in humans. Photodynamic therapy (PDT) using lasers has received increasing interest in the treatment of cancer and other diseases.

Photonic techniques have the potential for performing rapid in vitro screening of diseases and in vivo diagnosis on tissue without the need for sample excision and processing. Another advantage of photonics-based diagnoses is that the resulting information can be available in real time. In addition, because removal of tissue is not required for optical diagnoses, a more complete examination of the organ of interest can be achieved than with excisional biopsy or cytology. This chapter provides an overview of photonics technologies and their biomedical applications in the following areas: (1) in vitro diagnostics, (2) in vivo diagnostics and (3) photonics-based therapy.

C4.5.2 In vitro diagnostics

C4.5.2.1 DNA probes

Both microarrays and biochips use probes that consist of biological recognition systems, often called bioreceptors. Bioreceptors, which utilize a biochemical mechanism for recognition, are the key to specificity for biochip technologies [6]. They are responsible for binding the analyte of interest to the biochip sensors for the detection measurements. The biological probe may consist of different types of molecular species referred to as bioreceptors (e.g. an antibody, an enzyme, a protein or a nucleic acid) or a living biological system (e.g. cells, tissue or whole organisms). Biochips utilize various types of detection methods including: (a) optical measurements (i.e. luminescence, absorption, surface plasmon resonance, etc), (b) electrochemical and (c) mass-sensitive measurements (i.e. surface acoustic wave, microbalance, etc).

Nucleic acids have been widely used as bioreceptors for microarray and biochip technologies [14–21]. In DNA biochips, the biorecognition mechanism involves hybridization of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), which are the building blocks of genetics. The complementarity of adenine:thymine (A:T) and cytosine:guanine (C:G) pairing in DNA forms the basis for the specificity of biorecognition in DNA biochips. If the sequence of bases composing a certain part of the DNA molecule is known, then the complementary sequence can be synthesized and labelled with an optically detectable compound (e.g. a radioactive label or a fluorescent label). When unknown fragments of single-strand DNA samples, called the target, react (or hybridize) with the probes on the chip, double-strand DNA fragments are formed only when the target and the probe are complementary according to the base-pairing rule. When the targets contain more than one type of sample, each is labelled with a specific tag. The microarrays of probes on DNA biochips serve as reaction sites, each reaction site containing single strands of a specific sequence of a DNA fragment. The DNA fragments can either be short oligonucleotide (about 18–24) sequences or longer strands of complementary DNA (cDNA). Sample DNA can be amplified by the polymerase chain reaction (PCR), and a label (usually fluorescent) can be inserted during the PCR process. Finally, the sample is tested for hybridization to the microarray by detecting the presence of the attached labels.

Probes based on a synthetic biorecognition element, peptide nucleic acid (PNA), have been developed [17]. PNA is an artificial oligo amide that is capable of binding very strongly to complementary oligonucleotide sequences. Other advances include a relatively new type of spectral label for DNA probes based on surface-enhanced Raman scattering (SERS) detection [22–25]. The SERS gene probe technique has been developed for use in cancer diagnostics (figure C4.5.1) [25]. The development of a biosensor for DNA diagnostics using visible and near infrared (NIR) dyes has also been reported [26, 38].

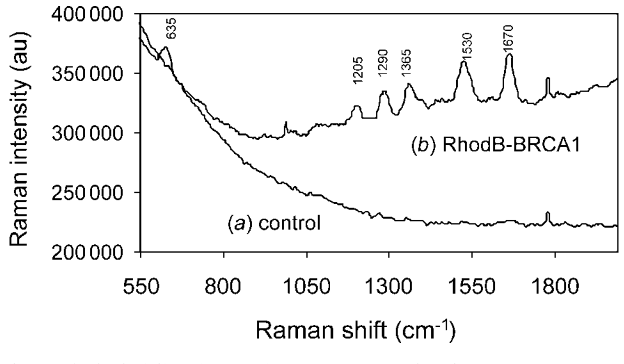


Figure C4.5.1. Principle of DNA hybridization using SERS probe for cancer gene diagnostics. Hybridization results for the BRCA1 breast cancer susceptibility gene: SERS spectrum on a silver island substrate of (*a*) the hybridization control and (*b*) the hybridized BRCA1 oligonucleotide labelled with Rhodamine B. The control used was the modified silver surfaces with self-assembled monolayer (SAMs) of alkylthiols subjected to all hybridization steps. (Taken from [25].)

C4.5.2.2 Biochips and microarrays

DNA biochips and microarrays can be grouped into two general classes: (i) DNA microarrays that consist of high-density probes on plate-based or gel-based substrates; (ii) integrated biochip devices that include DNA arrays and integrated circuit (IC) microchip sensors.

DNA microarrays

DNA microarrays generally consist of some type of substrate having microarrays of DNA probes. The substrates are thin plates made of silicon, glass, gel or a polymeric material such as plastic or nylon. These systems have high-density probe microarrays (10^3–10^5) and do not have integrated microsensor detection systems. These array plate biochips usually have separate detection systems that are relatively large (tabletop size) and are designed for laboratory-based research applications (figure C4.5.2). The large numbers of probes can be used to identify multiple biotargets with very high speed and high throughput by matching with different types of probes via hybridization. Therefore, array plates are very useful for gene discovery and drug discovery applications, which often require tens of thousands of assays on a single substrate.

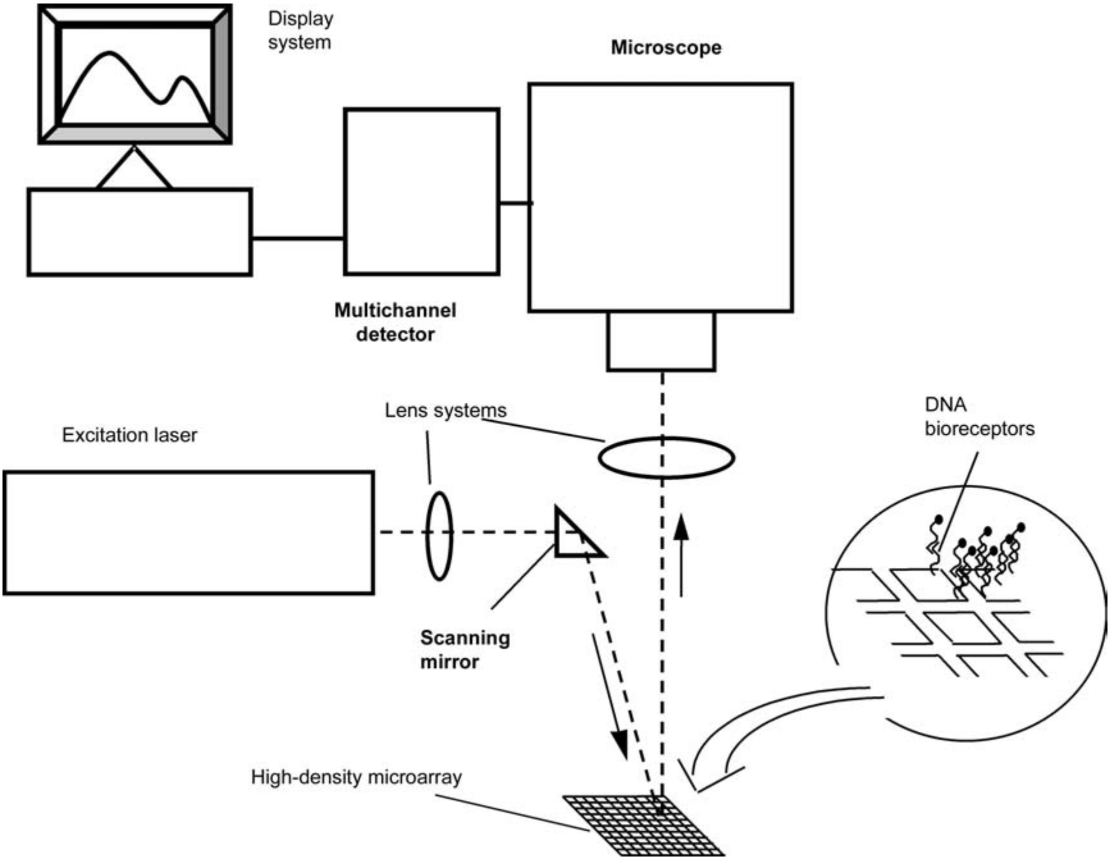

Display system
Microscope
Multichannel detector
Lens systems
DNA bioreceptors
Excitation laser
Scanning mirror
High-density microarray

Figure C4.5.2. Schematic diagram of microarray systems.

Various methods can be used to fabricate DNA microarrays. DNA arrays are typically produced using high-speed robotics on glass or nylon substrates. Oligonucleotide microarrays are fabricated either by in situ light-directed combinatorial synthesis that uses photographic masks for each chip [27] or by conventional synthesis followed by immobilization on glass substrates [2, 3, 28]. Activated DNA has been immobilized in aldehyde-containing polyacrylamide gel for use in manufacturing microarrays [29]. A maskless fabrication method of light-directed oligonucleotide microarrays using a digital microarray has been reported [30]. In this method, a maskless array synthesizer replaces the usual chrome mask with virtual masks generated on a computer, which are relayed to a digital microarray. Programmed chemical coupling cycles follow light exposure, and steps are repeated to achieve the desired pattern [30]. Modified chemistries should allow for other types of microarrays to be manufactured using photolithography. Alternative processes, such as ink jet [31] and spotting techniques, can also be used [29, 32].

DNA chip microarrays are externally connected to a photosensing system, which generally consists of a conventional detection device, such as a photomultiplier, or a charge-coupled device (CCD). Although the probes on the sampling platform are small, the entire device containing excitation laser sources and detection systems (often a confocal microscope system) is relatively large, e.g. tabletop size. These systems have demonstrated their usefulness in genomics research and analysis, but they are laboratory oriented.

Integrated biochips

Unlike laboratory-based microarray systems, integrated biochips also include an integrated circuit microsensor chip, which make these devices very portable, and inexpensive [6]. These devices generally have medium-density probe arrays (10–100 probes) and are most appropriate for field measurements or for medical diagnostics at the physician's office. Biochips that integrate conventional biotechnology with semiconductor processing, micro-electro-mechanical systems (MEMS), optoelectronics, and digital signal and image acquisition and processing, have received a great deal of interest.

The development of integrated biochips having a phototransistor integrated circuit (IC) microchip has been reported [7, 9–11, 13]. In addition, biochips with biofluidics systems have been developed. Figure C4.5.3 depicts a schematic diagram of such an integrated biochip having optical biosensor arrays on an integrated circuit. This multi-functional biochip has both DNA and antibodies as probes. The integrated circuit of the 16-channel sensing array of the biochip is shown in figure C4.5.4 [8]. In this biochip the sensors, amplifiers, discriminators and logic circuitry are all built onto the chip. In one biochip system, each of the sensing elements is composed of 220 individual phototransistor cells connected in parallel to improve the sensitivity of the instrument. An important element in the development of the integrated DNA biochip involves the design and development of an IC electro-optic system for the microchip detection elements using the complementary metal oxide semiconductor (CMOS) technology. With this technology, highly integrated biochips are made possible partly through the capability of fabricating multiple optical sensing elements and microelectronics on a single system.

C4.5.2.3 Medical applications of DNA biochips and microarrays

This section provides a brief overview of some applications of microarrays and biochips in a wide variety of medical areas including: (1) gene sequencing, mapping and expression analysis, (2) gene discovery and molecular profiling, (3) pharmacogenomics and toxicogenomics, (4) biomedical screening of genetic diseases, (4) public health protection and (5) medical diagnosis at the site-of-care.

Human expressed sequence tags (ESTs) in combination with high-density cDNA microarray techniques have been used to map the genome, to explore genes, and to profile downstream gene changes in a host of experiments. EST is a short strand of DNA that is a part of a cDNA molecule and can act as

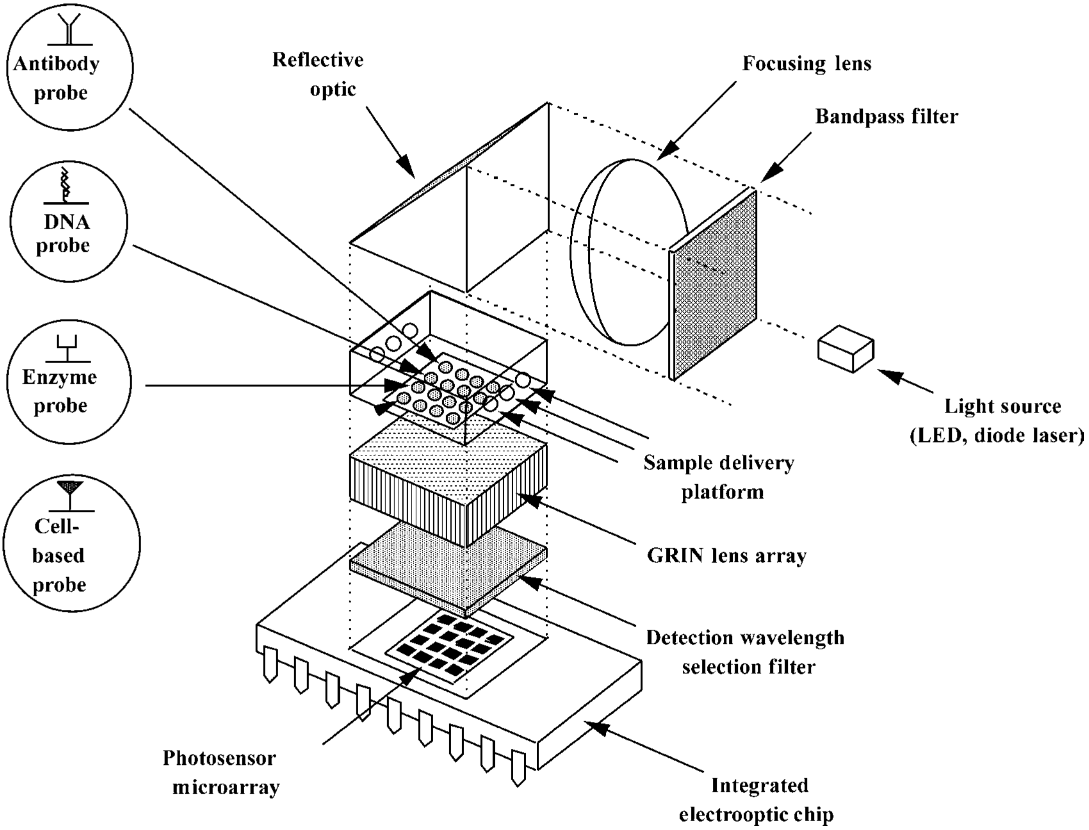

Antibody probe
DNA probe
Enzyme probe
Cell-based probe
Reflective optic
Focusing lens
Bandpass filter
Light source (LED, diode laser)
Sample delivery platform
GRIN lens array
Detection wavelength selection filter
Photosensor microarray
Integrated electrooptic chip

Figure C4.5.3. Schematic diagram of a multi-functional biochip having DNA and antibody probes.

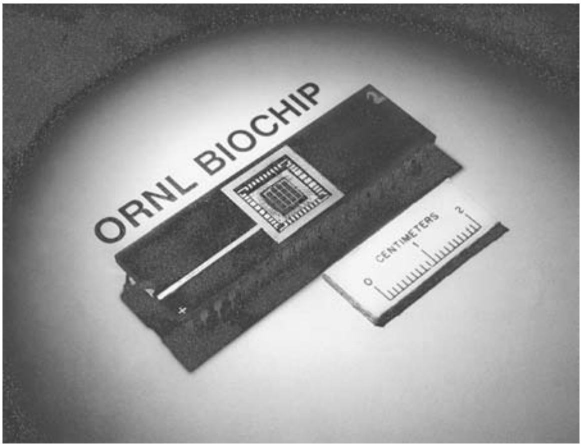

ORNL BIOCHIP
CENTIMETERS
0
1
2

Figure C4.5.4. Photograph of a CMOS integrated circuit system of a biochip (adapted from [8]).

the identifier of a gene. ESTs represent sequences expressed in an organism under particular conditions that are often used in locating and mapping genes. For instance, in lung cancer the effects of overexpression of a tumour suppressor—the phosphatase and tensin homology deleted on chromosome 10 (PTEN)—has been studied [33]. In prostate cancer, ESTs have been used for identification of several potential markers for prostate cancer [34]. High-density DNA microarrays have been used in the identification of sequence (gene/gene mutation) and determination of expression level (abundance) of genes. For the human genome, sequence information of human expressed sequence tag (EST) fragments is now available and has been used to evaluate gene expression in different types of cells or in different conditions and environments [34]. The power of the DNA biochip technology is the ability to perform a genome-wide expression profile of thousands of genes in one experiment.

Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome. Although more than 99% of human DNA sequences are the same across the population, variations in DNA sequence can have a major impact on how humans respond to disease; to such environmental insults as bacteria, viruses, toxins and chemicals; and to drugs and other therapies. Methods are being developed to detect different types of variation, particularly SNPs, which occur about once every 100–300 bases. SNPs provide powerful tools for a variety of medical genetic studies. A combination of gel-based sequencing and high-density microarrays DNA biochips was used in a large-scale survey for SNPs, where 2.3 megabases of human genomic DNA were examined [35].

Complex changes in patterns of gene expression are observed in the development and progression of cancer and the experimental reversal of tumorigenicity. Hence, microarrays of cDNA provide a powerful tool for studying these complex phenomena. For example, a high-density microarray of 1161 DNA elements was used to search for differences in gene expression associated with tumour suppression [4, 36]. High-density DNA array chips allow massively parallel gene expression and gene discovery studies and could provide researchers information on thousands of genes simultaneously. These arrays can be used to test an individual gene throughout all or part of its length for single-base variations from the gene's usual sequence. For example, DNA chips are commercially available to detect single-base variations in the human genes *BRCA1* and *BRCA2* (breast cancer related), *p53* (a tumour suppressor gene mutated in many forms of cancer), and *P450* (coding for a key liver enzyme system that metabolizes drugs).

Pharmacogenomics is a field aimed at establishing correlations between therapeutic responses to drugs and genetic profiles of patients. This information could lead to the development of new diagnostic procedures and therapeutic products that enable drugs to be prescribed selectively to patients for whom they will be effective and safe. The goal of toxicogenomics is to find correlations between toxic responses to toxicants and changes in the genetic profiles of the objects exposed to such toxicants. Here again, DNA biochips, which allow the monitoring of the expression levels of thousands of genes simultaneously, provide an important analytical tool to toxicogenomics.

DNA biochips could provide useful tools for screening genetic diseases. These screening procedures are often performed prenatally. DNA biochip technology could be a sensitive tool for the prenatal detection of genetic disorders. Genetic studies have identified a number of genes that must be mutated in order to induce cancer or promote the growth of malignant cells. Mutations in BRC1, a gene on chromosome 17, were found to be related to breast cancer. For example, a high percentage of women with a mutated BRC1 gene will develop breast cancer or have an increased risk for ovarian cancer. DNA biochips are available for the analysis of many human genes, including *BRCA1, BRCA2, p53* and *P450*. With the availability of the sequence data of the human genome, it is expected that similar biochips will soon be used for other tests.

Another important application of compact integrated biochip devices involves the detection of biological pathogens (e.g. bacteria and viruses) present in the environment and at occupational sites including hospitals and offices. DNA biochip technologies could offer a unique combination of performance capabilities and analytical features of merit not available in any other bioanalytical system currently available. Biochip devices that combine automated sample collection systems and multichannel-sensing capability will allow simultaneous detection of multiple pathogens present in complex environmental samples. In this application, the biochip technology could provide an important warning tool of exposure to pathogenic agents for use in human health protection and disease prevention.

At the physician's offices, portable, integrated biochip systems offer great advantages in terms of size, performance, fabrication, analysis and production cost due to their integrated optical sensing microchip. Figure C4.5.5 illustrates the detection of the cancer biomarker *p53* protein and the *Microbacterium tuberculosis* gene fragment using the multi-functional biochip [6]. The small sizes of the probes (microlitre to nanolitre) minimize sample requirement and reduce reagent and waste requirements. Highly integrated systems lead to a reduction in noise and an increase in signal due to the improved efficiency of sample collection and the reduction of interfaces. The capability of large-scale production using low-cost IC technology is another important advantage. The assembly process of various components is made simple by integration of several elements on a single chip. For medical applications, this cost advantage will allow the development of extremely low cost, disposable biochips that can be used for in-home medical diagnostics of diseases without the need of sending samples to a laboratory for analysis.

C4.5.3 In vivo diagnostics

Photonic techniques can provide a powerful means for detecting early neoplastic changes. Detection of early neoplastic changes is important because once invasive carcinoma and metastases have occurred, treatment is difficult. At present, excisional biopsy followed by histology is considered to be the 'gold standard' for the diagnosis of early neoplastic changes and carcinoma. In some cases, cytology rather than excisional biopsy is performed. These techniques are powerful diagnostic tools because they provide high-resolution spatial and morphological information of the cellular and subcellular structures of tissues. The use of staining and processing can enhance contrast and specificity of histopathology. However, both of these diagnostic procedures require physical removal of specimens followed by tissue

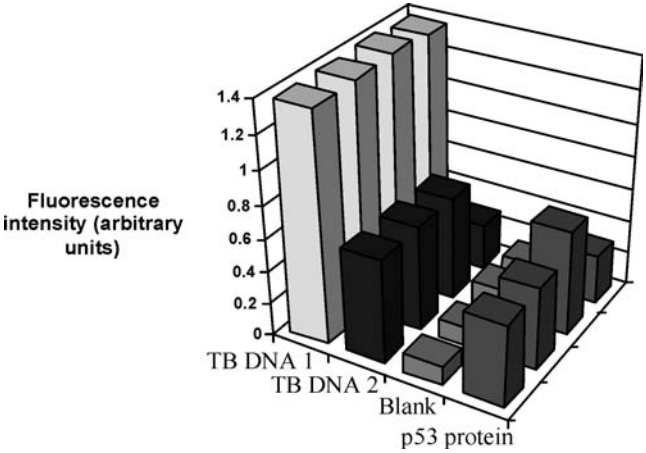


Figure C4.5.5. Simultaneous detection of the p53 protein (antibody probe) and the *Microbacterium tuberculosis* gene (DNA probe) using the multi-functional probe. (Taken from [6].)

processing in the laboratory. As such, these procedures incur a relatively high cost because specimen handling is required and, more importantly, diagnostic information is not available in real time. Moreover, in the context of detecting early neoplastic changes, both excisional biopsy and cytology can have unacceptable false negative rates often arising from sampling errors.

Photonic technologies have the potential capability of performing in situ diagnosis on tissue without the need for sample excision and processing. This diagnostic information can be available in real time. In addition, because removal of tissue is not required for optical diagnostics, a more complete examination of the organ of interest can be achieved with excisional biopsy or cytology.

Currently used optical diagnostic technologies can be broadly classified into two categories: (i) spectroscopic diagnostics and (ii) optical imaging. Spectroscopic diagnostics are used to measure some spectroscopic properties that are related to the molecular composition and structure of biochemical species in the tissue of interest. There are several spectroscopic methods that are used for optical diagnostics: fluorescence, elastic scattering, Raman (inelastic scattering), infrared absorption, etc. When a photon irradiates a molecule, it may be transmitted, absorbed or scattered. Different techniques arise from these different light–tissue interactions. These techniques include:

- absorption spectrometry;
- reflectance spectrometry;
- fluorescence spectrometry;
- raman spectrometry.

Each of these techniques, which have been studied for the purpose of tissue diagnosis with varying degrees of success, will be described in the following section.

C4.5.3.1 Fluorescence techniques

Fluorescence methods are important tools for medical diagnostics [37], and can be grouped into two main categories: (i) methods that detect endogenous fluorophores of tissues, often referred to as autofluorescence and (ii) methods that detect or use exogenous fluorophores or fluorophore precursors, such as 5-aminolevulinic acid (ALA). In other words, the fluorescence may originate from native fluorescent chromophores already present in the tissue or from an administered exogenous chromophore that has been specifically synthesized so that it targets specific tissue (e.g. dysplasia versus normal) or is activated by functional changes in the tissue. This section mainly discusses the autofluorescence method that involves detection of the fluorescence emission signal from the tissue itself, reflecting variations in the biochemistry of the tissue, as reflected in changes in fluorescent chromophores. The tissue is exposed to excitation light at some specific wavelength (typically near ultraviolet or visible), which excites tissue molecules, and induces fluorescence emission from these molecules; the emission spectrum (emission intensity versus wavelength) is then measured by varying the emission wavelength.

A number of investigators have investigated laser-induced fluorescence (LIF) as a method to discriminate tumours from normal tissues and diagnostic methods for lung can benefit from the extensive research conducted for detecting cancer in other organs. LIF spectrometry is a powerful technique that can potentially be used to noninvasively examine tissue fluorescence spectral signatures. Vo-Dinh and co-workers have developed a laser-induced fluorescence diagnostic procedure for in vivo detection of GI cancer that uses 410-nm laser light from a nitrogen pumped dye laser passed through a fibre-optic probe to excite the tissue [38–40]. After the tissue is excited, the resulting fluorescence emission is collected by the same fibre-optic probe and is recorded on an optical multichannel analyser in less than a second. Based upon the resulting fluorescence spectra, a diagnostic technique known as differential normalized fluorescence was employed to enhance the slight spectral differences between normal and malignant tissues [41, 42]. This technique greatly improves the accuracy of diagnosis, as compared to direct intensity measurements, because each spectrum is normalized with respect to its total integrated intensity and therefore becomes independent of the intensity factor. This in turn enhances small spectral features in weak fluorescence signals, making classification much easier. The sensitivity of the DNF method in classifying normal tissue and malignant tumours is 98% [39, 42]. The LIF methodology was also employed in clinical studies of over 100 patients, in which Barrett's mucosa without dysplasia was diagnosed with a specificity of 96% and high-grade dysplasia was diagnosed with a sensitivity of 90% [38, 40]. Figure C4.5.6 shows a schematic diagram of an LIF system for GI diagnostics [38–40].

The ability to distinguish between various types of tissues, in vivo, based upon multi-component analysis has been demonstrated [43]. Richards-Kortum and co-workers have used LIF, employing 337-nm excitation to differentiate in vivo cervical intraepithelial neoplasia (CIN), non-neoplastic abnormal, and normal cervical tissues from one another [44]. Laser-induced autofluorescence spectrometry has been investigated to detect colonic dysplasia in vivo using an excitation wavelength of 370 nm [1]. In a study of lung cancer, it was found that the sensitivity of the autofluorescence bronchoscopy was 86%, which is 50% better than conventional white-light bronchoscopy, for the detection of dysplasia and CIS [45, 46]. Like other cancers involving mucosal membranes, oral and laryngeal carcinomas have also been studied by autofluorescence. Fluorescence spectrometry has been used to differentiate normal tissue from dysplastic or cancerous tissue with a sensitivity of 90% and a specificity of 88% in a training set and a sensitivity of 100% and a specificity of 98% in a validation set [47].

An alternative approach to conventional fixed-excitation fluorescence is the synchronous luminescence (SL) method, which involves scanning both excitation and emission wavelength simultaneously while keeping a constant wavelength interval among them [48, 49]. This method has been developed for multi-component analysis and has been used to obtain fingerprints of real-life samples and for enhancing selectivity in the assay of complex systems. This SL procedure has been shown to simplify the emission spectrum and provides for greater selectivity when measuring the fluorescence or phosphorescence from mixtures of compounds. Spectral differences in SL emission profiles are related to the specific macromolecule(s) that differed between neoplastic and normal cells. The SL method has

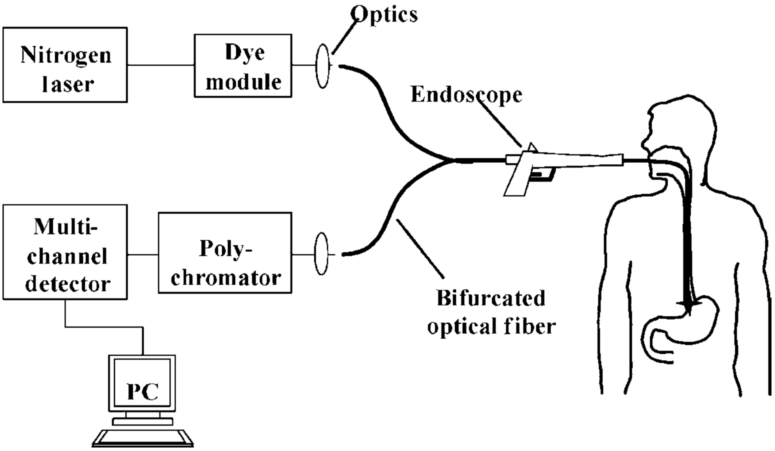


Figure C4.5.6. Schematic diagram of an laser-induced fluorescence system for cancer diagnostics.

been investigated as a tool for in vivo biomedical diagnostics [50], and a rapid screening method for monitoring DNA damage (DNA adduct metabolites) in animal studies, thus providing a technique for early cancer prescreening at the DNA level [51]. The SL technique has been shown to improve spectral selectivity in detecting normal and cancer cells for potential use in biomedical diagnostics [52], and to study the effect of tumour repressor gene [53].

C4.5.3.2 Elastic scattering techniques

The elastic scattering (ES) technique involves detection of the backscattering of a broad-band light source irradiating the tissue [54]. A spectrometer records the backscattered light at various wavelengths and produces a spectrum that is dependent on tissue structure as well as chromophore constituents. Tissue structure is the parameter that a pathologist assesses in performing diagnosis using histological examination. In general, the tissue is illuminated with the excitation light launched into a specific point location via an optical fibre and the scattered light is measured from a nearby location. Because light is strongly scattered in tissue, which is a highly diffusing medium, the mean free path for anisotropic scattering is several microns. Therefore, a stochastic process of many random steps generally describes the scattering path of photons. Calculations based on Monte Carlo methods have been used to investigate the photon scattering process. In some variations of this type of methodology, the optical transport properties of the tissue can be measured directly. The physical quantities measured in this diagnostic are the absorption and scattering properties of the tissue and/or the wavelength dependence of these properties. The ES technique has been developed for in vivo cancer diagnostics [55–57].

C4.5.3.3 Raman and infrared (IR) techniques

Raman scattering (RS) and IR absorption spectroscopies are vibrational techniques, which typically have high specificity because they provide vibrational information intimately related to molecular structures. Thus, the Raman technique can also be used for qualitative identification of biological compounds as well as for in vivo medical diagnostics [58]. The selection rules and relative intensities of IR and Raman peaks are not similar, so that RS and IR spectroscopies are often viewed as complementary. With IR spectrometry, the ever-present intense absorption bands of water (present in all biological samples), which overlap with most of the other tissue component spectra, could hamper IR spectrometry.

Raman spectrometry provides spectral information complementary to that obtained by fluorescence. The energy transitions of molecules are solely between the vibrational levels. When a photon is incident on a molecule, it may be transmitted, absorbed or scattered. Fluorescence arises from the emission of absorbed energy. Raman scattering arises from perturbations of the molecule that induces vibrational or rotational transitions. Only a limited number of biological molecules such as flavins, porphyrins and structural proteins (collagen and elastin) contribute to tissue fluorescence, most with overlapping, broadband emission. Most biological molecules are Raman active with fingerprint spectral characteristics; hence, vibrational spectrometry may provide another alternative for diagnosis of precancers and cancers. However, due to the intrinsically small Raman cross-section, the magnitude of the Raman scattering is typically small, making either high-illumination intensities or relatively long measurement times necessary in order to obtain good signal-to-noise ratios comparable to fluorescence techniques. Although any light source can be used, the intrinsically low intensity of Raman scattered radiation generally requires the use of laser radiation as the excitation source and, in fact, the use of Raman spectrometry has increased significantly because the development of high powered, continuous-wave gas ion lasers (e.g. argon and krypton ion lasers) in the late 1960s. An important feature of Raman spectrometry is the useful molecular information contained in the Raman spectra. For this reason, Raman spectrometry has potential for medical diagnostics and has been investigated for detection of cancer for many organs [59, 60, 61].

C4.5.3.4 Optical coherence tomography

Optical coherence tomography (OCT) is a technique that provides cross-sectional images of tissue in situ [62]. OCT is analogous to ultrasound B mode imaging that allows high-resolution cross-section imaging of the tissue microstructure. Instead of acoustic waves, OCT uses light and performs imaging by measuring the backscattered intensity of light from structures in tissues. In contrast to ultrasound, because the velocity of light is extremely high, the echo time delay of reflected light cannot be measured directly. Interferometric detection techniques must therefore be used [63]. An optical beam is focused into the tissue; the echo time delay of light reflected from the microstructure at different depths is measured by interferometry. Image information is obtained by performing repeated axial measurements at different transverse positions as the optical beam is scanned across the tissue. The resulting data yield a two-dimensional map of backscattering or reflectance from internal architectural morphology and cellular structure in the tissue. The axial resolution is $1-10\,\mu\mathrm{m}$, i.e. higher than any clinically available diagnostic modality. A device could be constructed with a fibre-optic probe, which can be incorporated into catheter-based or endoscopic systems.

The ‘histology-like’ images are obtained in real time with no need for excision, and thus this technique has potential for diagnosis of early neoplasia and surgical guidance. It is largely still a research technique, although it has been introduced into clinical applications in ophthalmology. OCT was originally developed and applied to tomographic diagnostics in ophthalmology. OCT can provide images of the retina with resolutions of $10\,\mu\mathrm{m}$. Imaging depth is limited by optical attenuation due to scattering and absorption. OCT has been applied in vitro to image arterial pathology where it can differentiate plaque morphology with a resolution superior to that obtained with ultrasound. OCT is a promising imaging technology because it can permit real-time and in situ visualization of tissue microstructure without the need to excisionally remove and process a specimen as in conventional biopsy and histopathology [63].

OCT can provide a diagnostic tool complementary to spectroscopic techniques and has great potential for in situ microscopic imaging of cellular attributes of malignancies and pre-cancers, which rivals that of a histopathologist examining a tissue biopsy specimen under a microscope. While cell staining techniques have not yet been designed to be visualized with this technique for use by the histopathologist, the orientation of the tissue within a three-dimensional matrix and the measurement of the sizes of cellular and sub-cellular elements in vivo can provide unique information and insights into the dysplastic and malignant process and can be linked to therapeutic procedures once a suspicious area is identified.

C4.5.3.5 Photon migration techniques

Photon migration techniques can be used to perform point measurements or imaging measurements in deep tissues (several centimetres) by using multiply scattered light. Great interest has been directed to time-resolved and phase-resolved in vivo spectrometry, particularly the measurement of the probability distribution of times for photons to travel from one point on a tissue surface to another [64]. The time delay of the photons coming out from another point on the surface a few centimetres away can be measured by using picosecond laser pulses launched into the tissue via optical fibres. The transit time is the optical path length divided by the speed of light in the tissue medium, and the attenuation of the signal intensity is due to the absorption within the tissue through which the light has migrated. The attenuation of the signal due to absorption depends on the molar extinction coefficient, the path length and the concentration of the absorbing species. Because the first two factors can be determined, these measurements might provide useful values of the concentration of the absorbing molecules. These techniques take advantage of the fact that light at near-infrared wavelengths is not highly absorbed by tissue and thus can penetrate several centimetres. Multiple scattering of the light degrades image information, so most of these techniques either focus on low-resolution imaging or functional assessment of tissue at low resolution. Because tissues are very strong diffusers of light, the mean free path between anisotropic scattering is on the order of 50 μm, and the light is almost completely randomized within 1 mm. Thus, for distances between two points in tissue greater than 1 mm, the migration of photons may be adequately described by a random walk of many steps. Various methods and instrumentation based on time–domain techniques [65, 66] and frequency–domain techniques [67–69] have been investigated to elucidate various aspects related to photon migration in tissues [70].

There has recently been great interest in theoretical studies to develop mathematical methods to reconstruct the image of tissues using photon migration techniques [69, 71, 72]. In the frequency–domain techniques, the incident light is intensity modulated, thus producing a photon number density, which oscillates in space and time. The resultant photon density waves scatter from tissue inhomogeneities (e.g. tumours in a tissue volume), and these scattered waves, if properly recorded, can be used to reconstruct the unknown inhomogeneity distribution into a two-dimensional image. The intensity-modulated photon-density waves obey a Helmholtz equation, and several investigators have attempted to exploit this fact to derive image reconstruction equations, using mathematical models analogous to that of ultrasonic diffraction tomography. Recent theoretical studies have been performed to derive a complete and explicit solution to this inverse problem of image reconstruction for diffraction tomography [72].

C4.5.3.6 Optical diagnostics using exogenous chromophores or precursors

Exogenous fluorophores are used for many reasons in clinical applications. The reason for using these compounds is often to provide a contrasting agent, which will make medical diagnoses easier, much like radionucleotides are sometimes used as contrast agents in circulation studies. However, due to the limited penetration of optical wavelengths into biological tissues, the most common type fluorescence analyses performed in vivo are cancer diagnoses of optically accessible tissues. It is for this reason that the majority clinical fluorescence diagnoses using exogenous fluorophores is in the area of cancer visualization. The most common of the exogenous fluorophores used for these studies are photosensitizers that have been developed for PDT treatments. These drugs generally exhibit strong fluorescence properties, and preferentially locate in malignant tissues.

Investigation of the optimal photosensitizer to use for various types of tumours is a continuous field of investigation. In one such study, fluorescence analyses of the HPD-type photosensitizer Photogem™ were performed in 22 patients with tumours of the lungs, larynx, skin; gastric and oesophageal carcinoma; and cancer of the gynaecological organs [73]. Monitoring and comparison of the results of ALA-induced PpIX tumour demarcation in chemically induced adenocarcinoma in the liver of rats and in an aggressive basal cell carcinoma in a patient were studied using LIF [74]. Investigation of Photofrin™-enhanced LIF differentiation of Barrett's metaplastic epithelium and oesophageal adenocarcinoma in five patients has been performed, following low dose intravenous injections ($0.35\,\mathrm{mg\,kg^{-1}}$ body weight). In this work, LIF measurements were performed on tissue specimens of normal mucosa, Barrett's epithelium and tumour tissue that had been treated with Photofrin™ [75]. Diagnosis of bladder cancer based upon the LIF of exogenous fluorophores has also been performed [76].

Another class of compounds that is beginning to be tested for in vivo tumour demarcation is fluorescently labelled antibodies. Based on previous experiments performed using nude mice, a feasibility study was performed to determine whether or not LIF of fluoresceinated monoclonal antibodies against carcinoembryonic antigens localized specifically in human carcinoma xenografts could be used for clinical colon cancer diagnoses. In this study, six patients with known primary colorectal carcinoma were given an intravenous injection of 4.5 or 9 mg of mouse–human chimeric anticarcinoembryonic antigen monoclonal antibody labelled with 0.10–0.28 mg of fluorescein. In addition, the monoclonal antibody was also labelled with 0.2 to 0.4 mCi of ^{125}I. Photodetection of the tumour was done ex vivo on surgically resected tissues from all six patients and in vivo by fluorescence rectosigmoidoscopy for one of the six patients. Fluorescence analyses revealed that the dye-labelled antibody localized preferentially in the tumour tissue at concentrations up to 0.059% of the injected dose per gram of tumour. This was 10 times greater than in the normal tissue, which exhibited a concentration of 0.006% of the injected dose per gram of normal mucosa. Such immunophotodiagnoses may prove very useful in the clinical setting for rapid tumor diagnoses in the colon and potentially other organs [77].

C4.5.4 Photonics-based therapy

C4.5.4.1 Laser-based therapy and surgery

Interstitial laser photocoagulation (ILP) is a laser-based procedure which uses an optical fibre (typically 0.2–0.4-mm core diameter) inserted directly into the target tissue (usually through a needle of about 18 gauge) so laser light is delivered to the tissue from the end of the fibre as a point source (bare tip fibre) or emitted from the end section of the fibre (diffuser fibre, where the diffuser section can be up to several centimetres long) [78]. The technique is sometimes referred to as laser interstitial thermal therapy (LITT) or interstitial laser thermotherapy (ILT). The energy is absorbed within the surrounding tissues causing local thermal necrosis. One or more fibres can be used. The ILT procedure generally delivers laser light (typically near-infrared) to a volume of tissue in order to gently heat the tissue to the point of denaturing proteins and DNA, generally at temperatures $\sim$60–90°C. This 'gentle' heating is designed to prevent tissue carbonization and vaporization, thus permitting the body's immune system to gradually remove the dead cells over a period of time. Various lasers have been tried but the most convenient are either a semiconductor diode or an Nd:YAG laser, at wavelengths between 805 and 1064 nm. By virtue, of the way that they are delivered, optical therapeutic techniques can often reduce the invasiveness of conventional surgical procedures, or enable new procedures that are not possible with conventional surgical tools. For example, interstitial laser thermotherapy has the potential to replace surgical excision in several applications, and has recently been approved for treatment of benign prostatic hyperplasia, as an outpatient procedure. In general, due to the availability of fibre-optic light delivery systems, laser therapeutics have the advantage that they can be performed in conjunction with many existing surgical or diagnostic modalities and tools such as endoscope, trocars, catheters, etc. With laser surgery, precision of therapeutic protocols can be enhanced because control of laser exposure parameters can be used to control the therapeutic process. For surgical applications, effects ranging from thermal cautery with haemostasis to precision ablation can be achieved.

The precise control of the wavelength as well as temporal and power parameters of laser therapeutic techniques can restrict the interaction to specific target areas of tissue. This important feature is beginning to be used in dermatology, where careful control of laser parameters permits selective destruction of specific loci in the skin, for example, in tattoo removal, treatment of port-wine stains and various cosmetic applications. For treatment of tumours of the breast, the procedure can be carried out under local anaesthetic and mild sedation. Up to four needles can be placed directly into the tumour either under ultrasound or MRI guidance. Optical fibres are then inserted through each needle. Once the fibres are in place the needle is withdrawn slightly so the bare ends of the fibres lie within the tumour [79].

C4.5.4.2 Photodynamic therapy (PDT)

An important application of optical technologies for site-selective therapy is the use of PDT [80, 81]. With PDT, a photoactive compound with some degree of selective affinity for cancerous tissue is administered topically, orally or intravenously. After a period of time (typically 6–48 h) the compound becomes concentrated selectively in areas of malignancy. The molecule is then photoactivated with light at a specific wavelength, producing singlet oxygen preferentially in malignant tissues. Although the exact mechanism of cell death is still under investigation, it has been shown that the presence of singlet oxygen in cells has a cytotoxic effect on the target cells.

Three critical elements are required for the initial photodynamic processes to occur: a drug that can be activated by light (a photosensitizer), light and oxygen. Interaction of light at the appropriate wavelength with a photosensitizer produces an excited triplet state photosensitizer that can interact with ground state oxygen via two different pathways, designated as Type I and Type II [82]. The Type II reaction that gives rise to singlet oxygen ($^1O_2$) is believed to be the dominant pathway because elimination of oxygen or scavenging of $^1O_2$ from the system essentially eliminates the cytocidal effects of PDT. Type I reactions, however, may become important under hypoxic conditions or where photosensitizers are highly concentrated. The highly reactive $^1O_2$ has a short lifetime ($< 0.04\ \mu s$) in the biological milieu and therefore a short radius of action ($< 0.02\ \mu m$). Consequently, $^1O_2$ mediated oxidative damage will occur in the immediate vicinity of the subcellular site of photosensitizer localization. Depending on photosensitizer pharmacokinetics, these sites can be varied and numerous, resulting in a large and complex array of cellular effects. Similarly, on a tissue level, tumour cells as well as various normal cells can take up photosensitizer, which, upon activation by light, can lead to effects upon such targets as the tumour cells proper, the tumour and normal microvasculature, and the inflammatory and immune host system. PDT effects on all these targets may influence each other, producing a plethora of responses; the relative importance of each has yet to be fully defined. Originally developed for treatment of various solid cancers, its applications have been expanded to include treatment of pre-cancerous conditions, e.g. actinic keratoses, high-grade dysplasia in Barrett's oesophagus, and noncancerous conditions, e.g. various eye diseases, e.g. age-related macular degeneration (AMD) [83]. PDT is also being investigated for applications in several clinical areas including skin cancer, bladder cancer, carcinoma of the GI tract and lung cancer [83].

An essential parameter in photodynamic therapy is the preferential uptake of the photosensitizer by neoplastic tissue. Photodynamic therapy is most beneficial if the light is delivered when the concentration of photosensitizer in tumour tissue is greater than that of adjacent normal tissue. Development of a noninvasive technique to estimate photosensitizer concentration in tissue, preferably in real time, is desirable. While fluorescence of photosensitizers may be used to localize tumours, it does not provide a quantitative measure of drug concentration in normal and tumour tissue.

C4.5.5 Conclusion

In conclusion, photonic technologies such as LIF, ES, RS and OCT have the ability to provide rapid in vivo diagnosis of diseases. Sensitivity and specificity in various organ systems will be the deciding factor on which technology is best for each organ or epithelium. Combination of these technologies may be important for improving sensitivity and specificity. The combination of these new optical technologies will be extremely useful for elucidating the underlying processes of various diseases and for increasing the accuracy, sensitivity and specificity of diagnostic and therapeutic features. One single technique may not be able to achieve these levels of accuracy, but a combination of techniques, such as optical endoscopy to screen large areas of epithelium followed by confirmation of abnormal areas by spectroscopic techniques. With these combined detection modalities, new information can be gathered, thus improving our understanding of the basic pathophysiology of not only cancers but also other, nonmalignant diseases, such as, benign structures, and inflammatory diseases whose mechanisms are currently poorly understood.

Another potentially important opportunity for optical technologies lies in the noninvasive measurement of the concentrations of various drugs and biological species in tissues. This capability would provide a variety of benefits in medical research. A specific example of such a benefit is the case of chemotherapy drugs used for the treatment of various cancers. While the therapeutic benefit is determined by the tissue concentration of the drug in the targeted organ or site, the only minimally invasive check available to the oncologist is to track the blood serum concentration and to assume a relationship between the amount of drugs at the target organs and the tissue concentration. More generally, the ability to track compound concentrations in tissue non-invasively would be a tremendous advantage. Optical methods could bypass much of the tedious and time-consuming trials that attempt to relate dosage to metabolic rates and target organ concentrations.

Finally, recent advances in biochip technologies, which combine biotechnology concepts and silicon fabrication technology, could lead to the development of inexpensive, portable or hand-held devices that can be used in the physician's office for rapid cancer screening by detecting genetic markers. Even if only one promise of optical diagnostic technologies is realized, that of reducing the number of required (often-random, costly and unnecessary) surgical biopsies, then this cost saving could be achieved with improved quality health care.

Acknowledgments

This work was sponsored by the National Institutes of Health (RO1 CA88787-01) and by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research, under contract DEAC05-00OR22725 with UT-Batelle, LLC.

References

[1] T Vo-Dinh, ed 2003 *Biomedical Photonics Handbook* (Boca Raton, FL: CRC Press)
[2] Schena M, Shalon D, Davis R W and Brown P O 1995 Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray *Science* **270** 467–470
[3] Blanchard A P and Hood L 1996 Sequence to array: probing the genome's secrets *Nat. Biotechnol.* **14** 1649–1649
[4] DeRisi J, Penland L, Brown P O, Bittner M L, Meltzer P S, Ray M, Chen Y D, Su Y A and Trent J M 1996 Use of a cDNA microarray to analyse gene expression patterns in human cancer *Nat. Genet.* **14** 457–460
[5] Wallace R W 1997 DNA on a chip: serving up the genome for diagnostics and research *Mol. Med. Today* **3** 384–389
[6] Vo-Dinh T 2003 Biochips and microarrays: tools for the new medicine *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)
[7] Vo-Dinh T, Alarie J P, Isola N, Landis D, Wintenberg A L and Ericson M N 1999 DNA biochip using a phototransistor integrated circuit *Anal. Chem.* **71** 358–363
[8] Vo-Dinh T 1998 Development of a DNA biochip: principles and applications *Sensors Actuators* **B51** 52–59
[9] Vo-Dinh T and Cullum B 2000 Biosensors and biochips: advances in biological and medical diagnostics *Fresenius J. Anal. Chem.* **366** 540–551
[10] Vo-Dinh T 1999 The multi-functional biochip *6th Annual Biochip Technologies: Chips for Hits'99* (Berkeley: Cambridge Health Institute)
[11] Vo-Dinh T, Griffin G D, Stokes D L and Wintenberg A L 2002 A multi-functional biochip for biomedical diagnostics and pathogen detection *Sensors Actuators* at press
[12] Vo-Dinh T and Askari M 2001 Micro arrays and biochips: applications and potential in genomics and proteomics *Current Genomics* **2** 399–415
[13] Vo-Dinh T, Cullum B M and Stokes D L 2001 Nanosensors and biochips: frontiers in biomolecular diagnostics *Sensors Actuators B—Chem.* **74** 2–11
[14] Wang J, Rivas G, Fernandes J R, Paz J L L, Jiang M and Waymire R 1998 Indicator-free electrochemical DNA hybridization biosensor *Anal. Chim. Acta* **375** 197–203
[15] Barker S L R, Kopelman R, Meyer T E and Cusanovich M A 1998 Fiber-optic nitric oxide-selective biosensors and nanosensors *Anal. Chem.* **70** 971–976

[16] Erdem A, Kerman K, Meric B, Akarca U S and Ozsoz M 1999 DNA electrochemical biosensor for the detection of short DNA sequences related to the hepatitis B virus *Electroanalysis* **11** 586–588

[17] Sawata S, Kai E, Ikebukuro K, Iida T, Honda T and Karube I 1999 Application of peptide nucleic acid to the direct detection of deoxyribonucleic acid amplified by polymerase chain reaction *Biosensors Bioelectron.* **14** 397–404

[18] Niemeyer C M, Boldt L, Ceyhan B and Blohm D 1999 DNA-directed immobilization: efficient, reversible, and site-selective surface binding of proteins by means of covalent DNA–streptavidin conjugates *Anal. Biochem.* **268** 54–63

[19] Niemeyer C M, Ceyhan B and Blohm D 1999 Functionalization of covalent DNA–streptavidin conjugates by means of biotinylated modulator components *Bioconjugate Chem.* **10** 708–719

[20] Marrazza G, Chianella I and Mascini M 1999 Disposable DNA electrochemical sensor for hybridization detection *Biosensors Bioelectron.* **14** 43–51

[21] Bardea A, Patolsky F, Dagan A and Willner I 1999 Sensing and amplification of oligonucleotide–DNA interactions by means of impedance spectroscopy: a route to a Tay-Sachs sensor *Chem. Commun.* **1** 21–22

[22] Vo-Dinh T, Stokes D L, Griffin G D, Volkan M, Kim U J and Simon M I 1999 Surface-enhanced Raman scattering (SERS) method and instrumentation for genomics and biomedical analysis *J. Raman Spectrosc.* **30** 785–793

[23] Vo-Dinh T 1998 Surface-enhanced Raman spectrometry using metallic nanostructures *Trac-Trends Anal. Chem.* **17** 557–582

[24] Vo-Dinh T, Houck K and Stokes D L 1994 Surface-enhanced Raman gene probes *Anal. Chem.* **66** 3379–3383

[25] Vo-Dinh T, Allain L and Stokes D L 2002 Cancer gene detection using surface-enhanced Raman scattering (SERS) *J. Raman Spectrosc.* **33** 511–516

[26] Vo-Dinh T, Isola N, Alarie J P, Landis D, Griffin G D and Allison S 1998 Development of a multiarray biosensor for DNA diagnostics *Instrum. Sci. Technol.* **26** 503–514

[27] Hacia J G, Woski S A, Fidanza J, Edgemon K, Hunt N, McGall G, Fodor S P A and Collins F S 1998 Enhanced high density oligonucleotide array-based sequence analysis using modified nucleoside triphosphates *Nucleic Acids Res.* **26** 4975–4982

[28] Blanchard G C, Taylor C G, Busey B R and Williamson M L 1990 Regeneration of immunosorbent surfaces used in clinical, industrial and environmental biosensors—role of covalent and noncovalent interactions *J. Immunol. Methods* **130** 263–275

[29] Proudnikov D, Timofeev E and Mirzabekov A 1998 Immobilization of DNA in polyacrylamide gel for the manufacture of DNA and DNA–oligonucleotide microchips *Anal. Biochem.* **259** 34–41

[30] Singh-Gasson S, Green R D, Yue Y J, Nelson C, Blattner F, Sussman M R and Cerrina F 1999 Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array *Nat. Biotechnol.* **17** 974–978

[31] McGall G, Labadie J, Brock P, Wallraff G, Nguyen T and Hinsberg W 1996 Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists *Proc. Natl Acad. Sci. USA* **93** 13555–13560

[32] Blanchard A P, Kaiser R J and Hood L E 1996 High-density oligonucleotide arrays *Biosensors Bioelectron.* **11** 687–690

[33] Hong T M, Yang P C, Peck K, Chen J J W, Yang S C, Chen Y C and Wu C W 2000 Profiling the downstream genes of tumor suppressor PTEN in lung cancer cells by complementary DNA microarray *Am. J. Respir. Cell Mol. Biol.* **23** 355–363

[34] Vaarala M H, Porvari K, Kyllonen A and Vihko P 2000 Differentially expressed genes in two LNCaP prostate cancer cell lines reflecting changes during prostate cancer progression *Lab. Invest.* **80** 1259–1268

[35] Wang D G, Fan J B, Siao C J, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris M S, Shen N P, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson T J, Lipshutz R, Chee M and Lander E S 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome *Science* **280** 1077–1082

[36] DeRisi J L, Iyer V R and Brown P O 1997 Exploring the metabolic and genetic control of gene expression on a genomic scale *Science* **278** 680–686

[37] Vo-Dinh T and Cullum B 2003 Fluorescence spectrometry in biomedical diagnostics *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[38] Panjehpour M, Overholt B F, Vo-Dinh T, Haggitt R C, Edwards D H and Buckley F P 1996 Endoscopic fluorescence detection of high-grade dysplasia in Barrett's esophagus *Gastroenterology* **111** 93–101

[39] Vo-Dinh T, Panjehpour M, Overholt B F, Farris C, Buckley F P and Sneed R 1995 In-vivo cancer-diagnosis of the esophagus using differential normalized fluorescence (Dnf) indexes *Lasers Surg. Med.* **16** 41–47

[40] Panjehpour M, Overholt B F, Schmidhammer J L, Farris C, Buckley P F and Vodinh T 1995 Spectroscopic diagnosis of esophageal cancer—new classification model, improved measurement system *Gastrointestinal Endosc.* **41** 577–581

[41] Vo-Dinh T, Panjehpour M, Overholt B F and Buckley P 1997 Laser-induced differential fluorescence for cancer diagnosis without biopsy *Appl. Spectrosc.* **51** 58–63

[42] Vo-Dinh T, Panjehpour M and Overholt B F 1998 Laser-induced fluorescence for esophageal cancer and dysplasia diagnosis *Advances in Optical Biopsy and Optical Mammography* (New York: New York Acad. Sciences) pp 116–122

[43] Sacks P G, Savage H E, Levine J, Kolli V R, Alfano R R and Schantz S P 1996 Native cellular fluorescence identifies terminal squamous differentiation of normal oral epithelial cells in culture: a potential chemoprevention biomarker *Cancer Lett.* **104** 171–181

[44] Ramanujam N, Mitchell M F, Mahadevan A, Warren S, Thomsen S, Silva E and Richards-Kortum R 1994 In vivo diagnosis of cervical intra epithelial neoplasia using 337-nm-excited laser-induced fluorescence *Proc. Natl Acad. Sci. USA* **91** (21) 10193–10197

[45] Lam S, Hung J Y C, Kennedy S M, Leriche J C, Vedal S, Nelems B, Macaulay C E and Palcic B 1992 Detection of dysplasia and carcinoma in situ by ratio fluorometry *Am. Rev. Respir. Dis.* **146** 1458–1461

[46] Zellweger M, Grosjean P, Goujon D, Monnier P, van den Bergh H and Wagnieres G 2001 In vivo autofluorescence spectrometry of human bronchial tissue to optimize the detection and imaging of early cancers *J. Biomed. Opt.* **6** 41–51

[47] Heintzelman D L, Utzinger U, Fuchs H, Zuluaga A, Gossage K, Gillenwater A M, Jacob R, Kemp B and Richards-Kortum R R 2000 Optimal excitation wavelengths for in vivo detection of oral neoplasia using fluorescence spectroscopy *Photochem. Photobiol.* **72** 103–113

[48] Vo-Dinh T 1978 Multicomponent analysis by synchronous luminescence spectrometry *Anal. Chem.* **50** 396–401

[49] Vo-Dinh T 1982 Synchronous luminescence spectroscopy—methodology and applicability *Appl. Spectrosc.* **36** 576–581

[50] Vo-Dinh T 2000 Principle of synchronous luminescence (SL) technique for biomedical diagnostics *Biomedical Diagnostics, Guidance and Surgical-Assist Systems II*, ed T G W S Vo-Dinh and D A Benaron (Bellingham, WA: SPIE)

[51] Uziel M, Ward R J and Vo-Dinh T 1987 Synchronous fluorescence measurement of bap metabolites in human and animal urine *Anal. Lett.* **20** 761–776

[52] Watts W E, Isola N R, Frazier D and Vo-Dinh T 1999 Differentiation of normal and neoplastic cells by synchronous fluorescence: rat liver epithelial and rat hepatoma cell models *Anal. Lett.* **32** 2583–2594

[53] Askari M, Miller G and Vo-Dinh T 2001 Synchronous luminescence: a simple technique for the analysis of hydrolysis activity of the fragile histidine triad protein *Biotechnol. Lett.* **23** 1697–1702

[54] Mourant J R and Bigio I J 2003 Elastic scattering spectrometry and diffuse reflectance *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[55] Mourant J R, Bigio I J, Boyer J, Conn R L, Johnson T and Shimada T 1995 Spectroscopic diagnosis of bladder cancer with elastic light scattering *Lasers Surg. Med.* **17** 350–357

[56] Bohorfoush A G 1996 Tissue spectrometry for gastrointestinal diseases *Endoscopy* **28** 372–380

[57] Bigio I J and Mourant J R 1997 Ultraviolet and visible spectroscopies for tissue diagnostics: fluorescence spectrometry and elastic-scattering spectrometry *Phys. Med. Biol.* **42** 803–814

[58] Mahadevan-Jansen A 2003 Raman spectroscpy: from benchtop to bedside *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[59] Mahadevan-Jansen A, Mitchell W F, Ramanujam N, Utzinger U and Richards-Kortum R 1998 Development of a fiber optic probe to measure NIR Raman spectra of cervical tissue in vivo *Photochem. Photobiol.* **68** 427–431

[60] Mahadevan-Jansen A, Mitchell M F, Ramanujam N, Malpica A, Thomsen S, Utzinger U and Richards-Kortum R 1998 Near-infrared Raman spectroscopy for in vitro detection of cervical precancers *Photochem. Photobiol.* **68** 123–132

[61] Puppels G J 2000, In vivo Raman spectroscopy *Microbeam Analysis 2000, Proceedings* pp 63–64.

[62] Fujimoto J G 2003 Optical coherence tomography imaging *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[63] Fujimoto J G 2001 Optical coherence tomography *C.R. Acad. Sci. Ser. Phys. Astrophys.* **2** 1099–1111

[64] Chance B (ed) 1989 *Photon Migration in Tissue* (New York: Plenum)

[65] Alfano R R, Demos S G, Galland P, Gayen S K, Guo Y C, Ho P P, Liang X, Liu F, Wang L, Wang Q Z and Wang W B 1998 Time-resolved and nonlinear optical imaging for medical applications *Advances in Optical Biopsy and Optical Mammography* (New York: New York Acad. Sciences) pp 14–28

[66] Beuthan J, Minet O and Muller G 1996 Quantitative optical biopsy of liver tissue ex vivo *IEEE J. Sel. Top. Quantum Electron.* **2** 906–913

[67] Lakowicz J R, Gryczynski I, Gryczynski Z and Johnson M L 2000 Background suppression in frequency–domain fluorometry *Anal. Biochem.* **277** 74–85

[68] Sevick-Muraca E M, Reynolds J S, Lee J, Hawrysz D, Thompson A B, Mayer R H, Roy R and Troy T L 1999 Fluorescence lifetime imaging of tissue volumes using near-infrared frequency domain photon migration *Photochem. Photobiol.* **69** 66S–66S

[69] Boas D A, Oleary M A, Chance B and Yodh A G 1997 Detection and characterization of optical inhomogeneities with diffuse photon density waves: a signal-to-noise analysis *Appl. Opt.* **36** 75–92

[70] Sevick-Muraka E M 2003 Near infrared fluorescence imaging and spectrometry in random media and tissues *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[71] Jacques S L, Ramanujam N, Vishnoi G, Choe R and Chance B 2000 Modeling photon transport in transabdominal fetal oximetry *J. Biomed. Opt.* **5** 277–282

[72] Norton S J and Vo-Dinh T 1998 Diffraction tomographic imaging with photon density waves: an explicit solution *J. Opt. Soc. Am. AMOpt. Image Sci. Vision* **15** 2670–2677

[73] Chissov V I, Solokov V V, Filonenko E V, Menenkov V D, Zharkova N N, Kozlov D N, Polivanov I N, Prokhorov A M, Pyhov R L and Smirnov V V 1995 Clinical fluorescent diagnosis of tumors using photosensitizer photogem *Khirurgiia (Mosk)* **5** 37–41

[74] Heyerdahl H, Wang I, Liu D L, Berg R, AnderssonEngels S, Peng Q, Moan J, Svanberg S and Svanberg K 1997 Pharmacokinetic studies on 5-aminolevulinic acid-induced protoporphyrin IX accumulation in tumours and normal tissues *Cancer Lett.* **112** 225–231

[75] von Holstein C S, Nilsson A M, Andersson-Engels, Willen R, Walther B and Svanberg K 1996 Detection of adenocarcinoma in Barrett's oesophagus by means of laser induced fluorescence *Gut* **39** (5) 711–716

[76] Baert L, Berg R, Vandamme B, Dhallewin M A, Johansson J, Svanberg K and Svanberg S 1993 Clinical fluorescence diagnosis of human bladder-carcinoma following low-dose photofrin injection *Urology* **41** 322–330

[77] Folli S, Wagnieres G, Pelegrin A, Calmes J M, Braichotte D, Buchegger F, Chalandon Y, Hardman N, Heusser C, Givel J C, Chapuis G, Chatelain A, Vandenbergh H and Mach J P 1992 Immunophotodiagnosis of colon carcinomas in patients injected with fluoresceinated chimeric antibodies against carcinoembryonic antigen *Proc. Natl Acad. Sci. USA* **89** 7973–7977

[78] Bown S and Bown S G 1983 Phototherapy of tumours *World J. Surg.* **7** 700–709

[79] Briggs G, Lee A and Bown S 2003 Laser treatment of breast tumors *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[80] Dougherty T J, Gomer C J, Henderson B W, Jori G, Kessel D, Korbelik M, Moan J and Peng Q 1998 Photodynamic therapy *J. Natl Cancer Inst.* **90** 889–905

[81] Dougherty T J 1985 Photodynamic Therapy *Clinics Chest Med.* **6** 219–236

[82] Henderson B and Gollnick S O 2003 Mechanistic principles of photodynamic therapy *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

[83] Dougherty T J and Levy G J 2003 Photodynamic therapy (PDT) and clinical applications *Biomedical Photonics Handbook*, ed T Vo-Dinh (Boca Raton, FL: CRC Press)

C5

The art of practical optoelectronic systems

Anthony E Smart

C5.1 Introduction

To create great paintings an artist not only needs a magnificent vision but must also know how to stretch canvas and mix paint. This chapter discusses practical skills for the design and implementation of optoelectronic systems. Since art has no agreed boundaries, technical depth may appear inconsistent, with items perhaps appearing trivial to one reader being useful insights for another, and vice versa. A specific optoelectronic system is chosen as an example to provide continuity while illustrating many techniques.

Departing from the more formal style of others, this chapter stresses practicalities, examples from painfully acquired experience, and methods of reducing risk. Ideas both familiar and less common are offered to permit the reader, at whatever level, to explore beyond the obvious, and to anticipate problems that may otherwise compromise success. While earlier chapters give depth and essential numerical specificity, practical arts are predominantly hints, recipes, and general guidelines. This chapter attempts to offer useful ideas to the beginner, while including sufficient subtleties to interest those with extensive experience.

The need for specifications and definitions of success (section C5.2), understanding requirements (section C5.3), and modelling (section C5.4), apply generally. Practicalities of design, planning, and organizational discipline (section C5.5) are important to the success of any complex optical and electronic endeavour. Rather than repeating quantitative details readily available elsewhere, emphasis is given to properties and things that go wrong at system design and component levels, with some ideas of how to avoid or recover from them (section C5.6). Later sections introduce testing (section C5.7), some effects of hostile environments (section C5.8), checklists (section C5.9), aphorisms (section C5.10), and a further reading list selected for its practicality.

C5.2 Specifications

Optoelectronic programmes can be put at risk by insufficient definition of the problem, the environment, the constraints, and/or the available tools. A dominant precursor to success in the art of optoelectronics is therefore to get a comprehensive written specification. Introducing our example we quote as table C5.1 the abbreviated target specification for an optical air data system (OADS), intended to replace the pneumatic system currently in the control loop of intrinsically unstable high performance aircraft, to allow operation closer to the limits of the flight envelope than otherwise possible. Outputs are aircraft speed and attitude with respect to the ambient air, whose temperature and pressure were also required.

The following sections exemplify methods used to meet these specifications, and how success may be defined. Table C5.2 shows a hierarchy of success, enhanced by proper specifications, top-down

Table C5.1. Optical air data system specifications.

Performance		
Speed (true air)	$15–1500\,\mathrm{m\,s^{-1}}$	$\pm 0.1\,\mathrm{m\,s^{-1}}$
Pitch (angle of attack)	+25 to −10°arc	±0.1°arc
Yaw (angle of sideslip)	±22.5°arc	±0.1°arc
Update rate	40 Hz	64 Hz preferred
Pressure (altitude)	−500 m to 16 km	±8 m (higher preferred)
Temperature (amb.)	−70 to 50°C	Unspecified (best)
Environment		
Altitude	Ground to 25 km	37 km preferred
Temperature	−70 to 140°C	Local ambient
Acceleration	−3 g to +12 g	Peak
Ambient light	Darkness to sun stare	Darkness to $1.5\,\mathrm{kW\,m^{-2}}$
Sensed particle radius	0.2 μm and larger	<0.12 μm preferred
Particle concentration	$<1\,\mathrm{ml^{-1}}$ to dense cloud	
Physical		
Throw	1.2 m	More preferred
Weight	<50 kg	Less preferred
Lifetime	100 000 h	MTBF
Final cost	<$100 k each	>1000 units
Power consumption	<1 kW	28 V ± 50 % (noisy)
Other		
Natural particles anywhere in global atmosphere; all weathers; eye-safe and stealthy		

Table C5.2. Hierarchy of success.

Ideal	All specifications are met
Adequate	Specifications are met and performance confirmed after acceptable modification
Incomplete	Not all specifications are met, but useful information and diagnostic data are recovered
Malfunction	The system produces no useful scientific results, but diagnostic data are recovered
Failure	No science is recovered and the system diagnostics do not indicate what went wrong
Catastrophe	The system destroys itself; we have no idea why
Disaster	The system destroys itself, causing collateral damage to the facilities and possibly personnel

conceptual design, and control of risk. Available components and circumstances may occasionally force bottom-up modifications.

C5.3 Requirements

The art of practical optoelectronics is about combining components, whose specifications we believe are plausible, into instruments and systems to solve problems for which we will be rewarded. At least four main factors are necessary; motivation, specifications, aptitude, and components. First, because optoelectronics is a challenging discipline, motivation must be sufficient, usually personal satisfaction or even financial gain. Second, we must demand explicit specifications and accept implicit boundary conditions, over both of which we have only limited control. Third, aptitude enhances the ability to create and configure an effective system. Frequently appearing as experience, it guides the designer and prevents unnecessary innovation. Fourth, the choice of components is based upon optimization of complex interrelationships of what is available, and what quoted claims are trustworthy. Possible consequences of the omission of one or more of these four factors are shown in figure C5.1. The practicality of any optoelectronic system also depends strongly on the control of risk, and the optimization of many complex interactions, not all of which may be fully understood nor can be wholly controlled.

Supplementary to these requirements, but at least as important for success, are time, money, and a champion. Even with ample resources things take time, but neither time nor money will be enough unless someone has made the personal commitment to succeed. Occasional digressions into areas less susceptible to scientific analysis are intended to help distinguish what is important from what is not, because much risk lies here.

Earlier chapters of this book give precise numerical facts about optical, electronic, and mechanical components, subsystems, and behaviours. Many are complex, interactive, and frequently counter-intuitive, typically more so in optics than in electronics. Familiarity with these subtleties is necessary, as is studying what others have done, and why, and whether these led to success or ingeniously redefined failure.

Almost all measurements and systems for which optoelectronics is conventionally used may be described as instrumentally assisted observation of a phenomenon. Ideally the phenomenon can be controlled, where it essentially becomes part of the system. It may however only be probed with negligible perturbation, as with the OADS example, or merely observed passively as in astronomy. The challenge is to create a specified instrument or experiment to enhance quantitative knowledge of the phenomenon, confirm existing wisdom, and perhaps suggest new ideas.

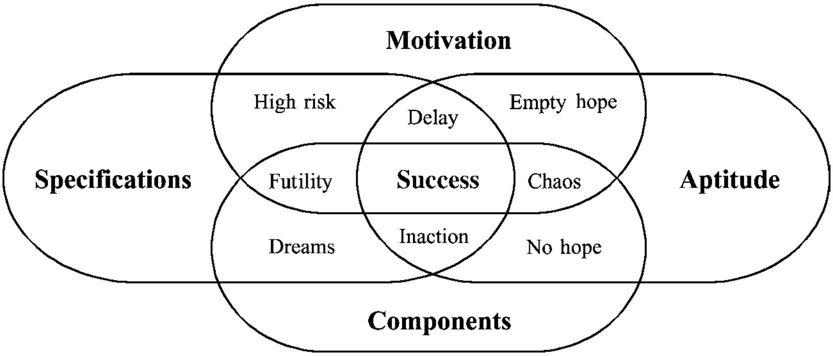


Figure C5.1. Venn diagram of success factors.

C5.4 Modelling

Making an optoelectronic system to the customer's satisfaction demands that specifications and requirements are understood. An initial approach may be to construct a mental model of a possible system, from concept through design, assembly, testing, operation, and interpretation of the output. Table C5.3 lists eight major functions of the initial thought experiment. At each design level new tasks are spun off, while questions are asked and answered to permit continuation of the planning and design sequence.

Eliminating or solving problems uncovered while addressing these items frequently redirects attention. For example, maintaining adequate temperature stability for OADS originally seemed impossible. Combining air circulation and heat pipes offered a possible if inelegant solution permitting design progress to continue, and although this might have been acceptable, the final choice to circulate a temperature stabilized liquid through a sealed labyrinth held temperatures with 0.5°C, maintaining optical alignment and electronic performance. Eliminating such show-stoppers as early as possible reduces redesigns, which get more difficult and expensive the later their necessity is accepted. The thinking process implicit in table C5.3 depends strongly upon feedback to achieve performance satisfactory to the customer and consistent with available resources.

Figure C5.2 maps a possible design sequence, where arrows represent potential information flow upon which changes may be contingent. The starting point is the customer's specification. If this appears infeasible or even unreasonably demanding, for example by violating the laws of physics or similarly extreme constraints, early and amicable negotiations can yield great advantages in simplicity, cost, and timeliness—less commonly in performance. Quantitative examination of the proposed physics and phenomenology ① may confirm acceptability or demand changes. For OADS, the three-dimensional velocity was calculated from the measured transit times of naturally occurring particles crossing each of three inclined pairs of parallel light sheets projected 1.2 m from the optical unit looking sideways from the aircraft. Pressure, altitude and temperature were obtained from simultaneous pulsed measurements of total elastic scattering and fluorescence of molecular oxygen over a common volume of air outside the boundary layer of the vehicle. Simple calculations ② with tests for scientific plausibility and technical capability can confirm that nothing silly is being planned ③. Plausible and self-consistent numerical criteria, which for OADS were difficult but not infeasible, may then allow a preliminary design ④. The projection of 8 mm light sheets with a near diffraction limited full width to e^{-2} intensity of 70 μm, separated by 11.2 mm, looked reasonable, but required transit time measurements accurate to 10^{-4}, challenging but not impossible—see below. External and internal interfaces with a coherent logical and structural architecture (figure C5.3), iterate to a self-consistent set of available parts ④, again. Clear and unambiguous names for each module help to avoid confusion among participants. Defined hardware and software modules must be isomorphic, with each other and also with the necessary functions and interfaces to allow testing and possible remediation. Within this structure an ideal

Table C5.3. Functions of a model.

1.	Quantification of the physics and phenomenology
2.	Verification of compatibility of specifications and constraints
3.	Understanding quantitative interactions between parameters
4.	Optimization of architecture, design configuration, and components
5.	Prediction of ideal performance possible with the planned design
6.	Prediction of performance deterioration when implementation is not ideal
7.	Diagnosis of anomalies anticipated during construction, testing, and use
8.	Reduction of risk and enhancement of confidence

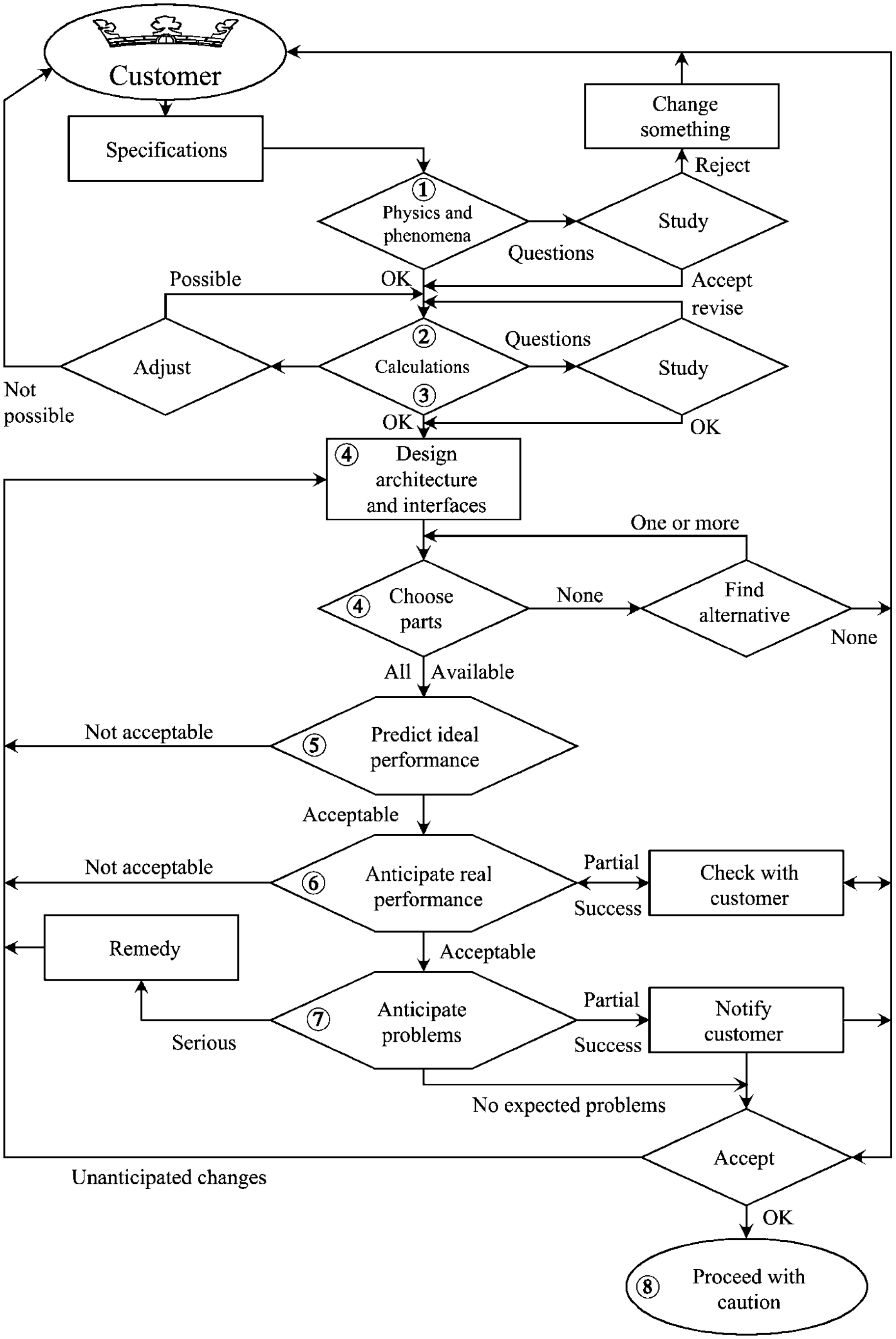

Customer
Specifications
Change something
Reject
① Physics and phenomena
Study
Questions
OK
Accept revise
Possible
② Calculations ③
Questions
Study
Adjust
Not possible
OK
OK
④ Design architecture and interfaces
One or more
④ Choose parts
None
Find alternative
None
All Available
⑤ Predict ideal performance
Not acceptable
Acceptable
⑥ Anticipate real performance
Not acceptable
Partial
Success
Check with customer
Remedy
Acceptable
⑦ Anticipate problems
Serious
Partial
Success
Notify customer
No expected problems
Accept
Unanticipated changes
OK
⑧ Proceed with caution

Figure C5.2. Operation of design model.

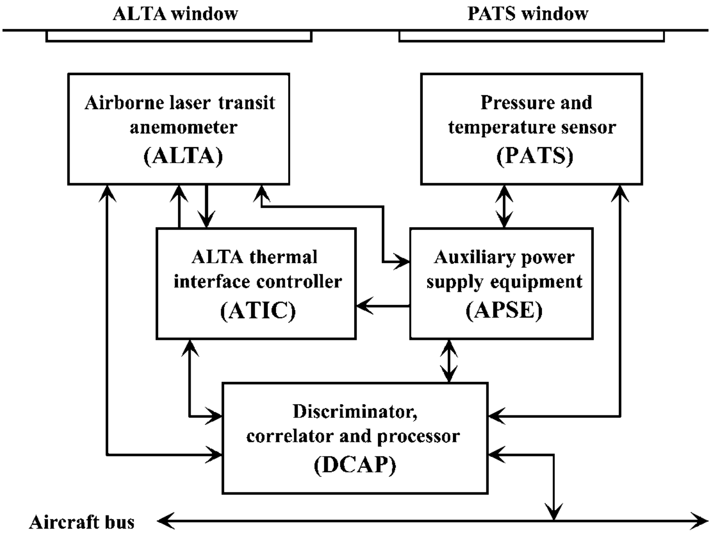

ALTA window
PATS window
Airborne laser transit anemometer (ALTA)
Pressure and temperature sensor (PATS)
ALTA thermal interface controller (ATIC)
Auxiliary power supply equipment (APSE)
Discriminator, correlator and processor (DCAP)
Aircraft bus

Figure C5.3. OADS system components.

performance can be specified ⑤. Independent knowledge, measurements, and/or testing allow consideration of effects of ageing, deterioration during use, or even failure of parts or subsystems to anticipate likely real performance ⑥. In critical applications where single point failure cannot be tolerated, initial choices for system architecture may require adaptation. Expanding the rigour of preliminary thoughts and testing at this level may show detrimental problems ⑦, requiring more serious remediation. For OADS, the pressure and temperature system (PATS) technology was changed from the originally planned differential absorption to Schumann–Runge fluorescence of O_2 because of installation constraints. However tedious and expensive preliminary modelling might appear, it is incomparably cheaper than later fixes, even should these prove possible ⑧. Design errors are better corrected on paper than in hardware components or computer code. Three outcomes from this exercise may be (1) no obvious problems, (2) problems with reasonable remedies, or (3) problems requiring negotiation with, or at least reporting to, the customer. Such a sequence of increasingly demanding analytical processes for OADS identified five major risks: (1) Would there be enough particles everywhere in the atmosphere? (2) Is the velocity measurement geometry acceptable? (3) Can the system be calibrated? (4) Is everything stable enough? (5) Would PATS work? In the event enough particles were found everywhere flown, the velocity measurements were sufficiently accurate, more so than systems used to calibrate them, stability was just acceptable, and PATS performance was marginal. Ironically the major risk was not technical, and the programme, of which OADS was but a tiny fraction, was cancelled by the customer's customer for wholly unconnected reasons.

However, rather than insisting that everything be modelled it is often more sensible to evaluate where effort is most wisely spent in terms of the overall requirements of the customer. Run-away costs, or late delivery, can be just as serious as an inadequate instrument.

C5.5 Design

Design, like art, is a creative process involving flashes of insight. However, the first practical step in building a cathedral is pounding rocks into a ditch, and in this sense the preparation of foundations is an important, if a less apparently rewarding precursor of more detailed design work. Using appropriate tools

Table C5.4. Tools and purposes.

1.	Riding the ray	Analyse all optical components
2.	Assembly methods	Provide adequately sensitive adjustments and criteria
3.	Calibration	Build trust in the planned measurements
4.	Operation	Use and experience the virtual system
5.	Risk reduction	Consider the unknowns, and the unknown unknowns
6.	Signal estimation	Predict signal, noise, accuracy, precision, and errors
7.	Information retrieval	Simulate signal processing and information recovery
8.	Interfaces	Estimate power, environment, and communication needs
9.	Management	Track status, changes, costs, and consequences

for design planning amply repays the effort; examples follow. A more colloquial style is adopted here to lighten the overwhelming complexity that can be engendered by rigorously performing the following tasks for which a fairly logical sequence is established below, and followed later in section C5.6.

A top-down design methodology helps to assure that specifications can and will be met. Important aspects of the system, although initially existing only in imagination, can be solidified into pathways along which the physical construction can later proceed. Nine ideas introduced in table C5.4 are expanded in the following paragraphs. General programme management and design activities must often be augmented for specific systems by special software packages for optical design, signal analysis optimization, computer aided design (CAD), and finite element analysis (FEA) for mechanical and thermal optimization.

C5.5.1 Riding the ray

Riding the ray tracks information-bearing photons, later transformed into electrons, through successive components. Examining the effects of each component, including the phenomenon, from generation of photons to the recovery of the numerical and graphical representations of the final measurement, can explore the properties of many devices in the system. This approach is particularly useful for examining the effects of aberrations, polarization, spatial and temporal coherence, optical power budget, pointing stability, alignment criteria, and various other component-related properties. Where there is more than one optical path, each may be analysed independently, or in various combinations. This approach is the essence of ray tracing, estimating aberrations and ghosts. It may also show unanticipated effects.

C5.5.2 Assembly methods

In addition to describing the structure, design must specify how the system may be constructed, how it will be aligned, and the alignment maintained. For example, mechanical support may be (1) fixed by design and manufacture, (2) initially adjustable over a specified range, or (3) dynamically compensated during operation. In mass production, where adjustments are neither necessary nor permitted, functionality may be compromised to accommodate manufacturing tolerances. A more versatile design avoids such performance reduction at the expense of increased complexity. The extreme is to have everything adjustable, with ranges and sensitivities sufficient to accommodate almost any manufacturing imperfection. This can be a dangerous and unwise choice. A better solution optimizes reasonably tight manufacturing tolerances with a few well-chosen adjustments, whose ranges are best limited to the smallest necessary. Each adjustment should be independent and appropriately sensitive. After meeting a clear criterion of acceptability, each should be locked, and not be susceptible to residual

Table C5.5. Adjustments and alignments.

Available adjustments
- Provide independent control for each variable
- Arrange suitable sensitivity of adjustment
- Assure that every adjustment may be rigidly locked

Alignment criteria
- Agree a definition of acceptable alignment
- Specify a quantitative criterion for its achievement

Completion
- Confirm that the alignment criterion is finally met
- Mark the final position and document all activities and observations

creep. If such stability is impossible, perhaps because of changing external conditions, then adjustments may be driven by a closed-loop feedback system controlled by a pre-specified algorithm to an established alignment or performance criterion. The most likely candidate for real time control is temperature stabilization, but other effects may be similarly mitigated.

Table C5.5 summarizes guidelines for mounting and aligning the chosen components, with specific techniques given later (section C5.6.3). For any particular system, some line items may appear either obvious or too detailed. It is important to know which, and assign effort intentionally, rather than by crisis intervention.

Every intentional, or even accidental, variable should be reviewed in the light of table C5.5. Listing all variables allows the definition of acceptability criteria for each alignment. Early attention to surprises that often emerge from this exercise greatly enhances assurance of success.

C5.5.3 Calibration

Although calibration is often necessary to assure the desired accuracy of measurements, it may be simplified by careful manufacture, by ingenious choice of parameters, or by specifying measurements of more stable output quantities, such as time and frequency, rather than intensity or voltage. If possible the reported measurement should be independent of quantities that are difficult to measure or guarantee constant, such as luminance, wavefront flatness, stray light, noise level, and others. Table C5.6 suggests a sequence of informal questions to which proper answers can improve confidence.

Table C5.6. Calibration and characterization.

Do the numerical results mean what they seem to mean?
What precision can be expected, or achieved?
What accuracy can be expected, or achieved?
What evidence links the current results to previously known results?
Can errors be bounded using comparisons of known and measured data?
Do the measurements depend exclusively on the phenomenon of interest?
If not, what other effects must be allowed for?
Can these effects be measured and the data corrected?
If not, can the errors be otherwise bounded or some compensation performed?

Calibration implies that the system is both stable and well characterized and that its initial behaviour is understood and close to what is intended. Residual disparities are assigned numerical values to correct or compensate measurements, possibly even while the instrument is operating. Implicit in calibration is the planning of methods to test the system for efficacy, performance, reliability, constancy, and other properties discussed in section C5.7.

C5.5.4 Operation

Anticipating what instructions may be necessary and what responses are expected from an operator is desirable. A proper balance between automatic operation and user intervention can accommodate unforeseen situations, without making routine tasks so complex that they invite errors. Routine operation should be automatic, perhaps run by pre-written scripts, requiring user intervention only if something unexpected arises. If results or monitored quantities differ from expectations, the operator can request additional diagnostic displays, pre-empt a script in favour of a new one, make changes to existing commands, or take over progressively more complete control of the system. This hierarchy of increasingly operator-intensive actions must be fairly robust to operator errors induced by fatigue or stress caused for example by something unexpected. Even if the operator must assume complete control, with increased risk of mistakes, on-line error checking can prevent incompatible instructions.

C5.5.5 Risk reduction

Table C5.7 lists five main areas of risk, although optoelectronics carries special types of risk with each element potentially leading to further questions. Experience helps identify critical areas, and often a barely remembered thought about some device or approach can intimate something that has perhaps been neglected, at future peril.

One great risk may be not staying aware of what is wanted, perhaps by misunderstanding the customer's original desire or intention, or even a change of priorities during design and/or development. Sharing risk with the customer by frequent timely exchanges of current status information, thoughts, feelings, concerns, intuitions, and expectations enhances the probability of a mutually successful outcome. Rarely will a customer become impatient with such attention, but neglect of personal relationships and communication is a sure way to fail. Of particular hazard in optoelectronics is sub-optimization, achieving flawless performance in areas with which one is most familiar, while less well understood areas receive less attention. The following paragraphs address areas of risk not always stressed in optoelectronics texts, but critical nevertheless.

Ideal and real component behaviours may differ. Buying or borrowing critical parts for early testing increases confidence. Design reviews are vital and once a design is accepted, changes should be

Table C5.7. Risk areas.

Customer	Are there enough timely, frequent, and honest exchanges?
Resources	Are planned schedule, funding, personnel, and expertise available?
Technical	Is the physics right to show the desired effects?
	Will the experiment design allow proper measurements?
	Is the system implementation acceptable in all areas?
Implementation	Can what is designed be built?
	Can adequacy be confirmed by well-specified testing?
Deployment	Will it stay that way during all conditions of use?

evolutionary rather than revolutionary, with ripple-through effects reduced by the early definition of modularity and interfaces.

Testing to limit functional risk can only be complete when physical hardware exists, too late for conceptual and design changes without great inconvenience. However, components, software and hardware modules, and supporting structures can be independently tested as soon as they are available. To be effective, testing requires a clear definition of what is expected, a valid criterion for its manifestation, and quantitative documentation of how and to what extent this is observed.

Robustness to external conditions, misuse, and misunderstanding is necessary once control of the system passes to a user who may typically have different attitudes and capabilities from those of the designer. The term foolproof is probably optimistic—fools are more creative than can be imagined.

Complexity increases the technical risk of not performing as wished, or of doing inadvertently something neither expected nor understood. Having fewer parts increases the chance of success. The least trustworthy components are complex or nonstandard parts for which little prior experience exists. Using common commercially available parts where possible, even with some slight performance compromise, is often recommended. Although this might apply particularly to optics and/or mounts, commercial mounting assemblies for optics may be intended for environments different from the current need, and some commitment to modification, such as adding locking screws, must be accepted. An almost trivial example of simplification is the specification of screw sizes. As far as practicable, the screws should be the same type and size; that way the ability to find the right wrench is much enhanced, especially if you buy many of that size. The ideal of tool-less assembly and maintenance is rarely achievable in optoelectronics.

In optics, common sense is often corrected by experience, and optical designs that seem acceptable on paper, or even in the laboratory, can fail in use in various subtle ways. Time and effort devoted to anticipating failure modes is better directed to the more likely ones, identified seemingly only by painfully acquired experience. Here, design for graceful degradation can permit remedial action before complete failure. Time invested with suppliers before committing to final design, parts, or assembly methods is never wasted.

Although durability is not always a direct concern, many optoelectronic systems are so diverse and complex that it is well to build them as though they must last forever, or at least for much longer than the specification demands. Three approaches may be useful. Design the system to tolerate all anticipated conditions. Preserve the system from unplanned conditions that could cause deterioration. Arrange continuous or frequent monitoring to assure that no relevant parameter is, or has been, outside its permissible range.

An optoelectronic system should be robust to any change in conditions within its specification. This has sometimes been called hardening, which is commonly used with three meanings, the robustness of (1) the design, (2) the resultant instrument, and (3) the ability to operate without serious deterioration in radiation environments. Even allegedly benign conditions can have hazards. From experience, the author expects to lose one or two digital camera charge coupled device (CCD) pixels to radiation damage on every long-haul commercial airline flight.

Although rarely in the specifications, a robust design not only establishes initial operation, but also accommodates inevitable later changes. Design quality is enhanced by having experienced people review it for potential weaknesses before commitment to procurement and manufacture. Reviewers are best selected from those who are sceptical that it can be made to work at all, and are looking for justification of their prejudice. Heeding their observations and improving the objects of their concern improves the robustness of the design to many unquantifiable properties of real life. Four aspects merit review: (1) general concept, (2) interface and layout drawings, (3) design frozen prior to commitment to hardware, and (4) software responsible for the operational experience of the user. Once testing is begun, the only permitted design changes should be those demanded by problems uncovered by or during that testing.

C5.5.6 Signal estimation

Quantitative estimates of expected system properties can be obtained by representing each component by a cell on a spreadsheet, entering a plausible estimate of the power or signal properties at that point, from source to detection, allowing for all the plausible or possible effects. Following columns contain best- and worst-case approximations that may raise confidence, suggest remedial action, or demand redesign. This technique can identify potential problems arising from many sources, from stray light to amplifiers or subsequent processing. Commonly, modern optoelectronic systems make at least two sequential transitions: optical radiation to analogue electrical signals, and digitization of the analogue signal. Unless the system is based upon individual photon detections that are in some sense already digital, converting from analogue to digital transmission as close to the detector as practicable minimizes further increase of noise from parasitic and transfer effects.

Optical noise sources

However good, detectors and electronics can never recover properties of the optical signal lost or diluted by unintended effects. Examples vary from a mismatched aperture to windows that deteriorate or get dirty, but can be improved by care and attention to component choice, environment, and operating methods. Stray light from various internal and/or external sources can dilute the optical signal. Most commercially available lens and optical design programs have the facility to analyse ghosts, which although reducible by multi-layer anti-reflection coatings, may still be detrimental in laser-based systems. With high power pulsed lasers even ghost reflections can focus enough of the beam to cause ionization along the optical path or permanent damage if the focus is near or within a component. Cleanliness is always essential. While superpolishing is sometimes recommended, it is expensive and parts require special handling.

Internal flare can be improved by machining fine annular grooves with a matt black coating on all enclosing barrels. Conventional optical stops may be augmented by a few well-placed discs with internal knife-edges to reduce stray light. Despite potential external attractiveness, shiny or polished surfaces should be avoided internally, substituting surface finishes such as black anodizing for matt aluminium, and various other oxidation processes for materials such as brass or steel. Black plastic shrouds also help reduce reflections and flare, and enclosing the light path often suppresses convection, an occasional source of image degradation where density gradients occur. Carbon-based matt black paints can be useful and cobalt oxide pigments survive and stay black at high temperatures.

The advent of lasers as illumination sources has introduced the phenomenon of cavity feedback, which is less rare than its cause might suggest. Even a tiny percentage of light fed back coherently into the laser cavity can destabilize cavity gain and hence the output intensity or mode shape. Even a minutely fluctuating feedback from optical components can be significantly amplified in the resonator to become a serious instability. Sensitivity to mechanical or thermal instabilities becomes more serious as the instrument is better aligned, or contains surfaces conjugate with the laser output face. For example, when launching the laser beam into a single-mode fibre, the best match of numerical aperture, focus, and centring is also the most likely condition for laser destabilization. Tilting the input face sometime helps, but wedging is better. The opposite end of the single-mode fibre is also conjugate, and if the fibre is shorter than the laser coherence length, or a near multiple of the beat length between close modes if more than one is stimulated, then this also will make the laser flicker.

Detection and noise

Currently the only directly detectable property of an electromagnetic field at optical frequencies is its intensity, although techniques such as coherent mixing can visualize phase, and spectral characteristics can be derived via various dispersive or nondispersive optical processes. Square-law detectors measure light intensity in one of three regimes: (1) low intensity, where quantum properties prevail, (2) high light levels, with classically continuous intensity levels, and (3) a 'grey' area between, where observed behaviour depends on the power level and detector type. Noise effects differ significantly between regimes, easily estimated by noting that a green photon has energy of about $4e^{-19}$ J, giving about $2.5e^{18}$ photons s^{-1} in a 1 W beam.

Detectors capable of resolving individual photons often show defects such as after-pulsing, fluctuating sensitivity, and photoelectron pulse pile-up, which can distort the statistics upon which information depends. If these effects are reduced to be negligible, the detector output at sufficiently low light levels, and/or high bandwidths, consists of a series of amplified photoelectron pulses. Even if many photons are missed because of reduced quantum efficiency, the statistics can remain an unbiased estimate of incident intensity, usually as Poisson distributed arrival times. Noise sources in the detector itself, the following amplifier, and its resistive load, usually become significant only in the classical regime. Certain types of detector, such as photomultipliers (PMTs) and avalanche photodiodes (APDs) in the Geiger mode, have enough internal gain to produce a discriminable electrical pulse from each released photoelectron. The statistical fluctuation, commonly called shot noise, which increases as the square root of the optical signal power, is an effect of quantization of the optical field, and thus is not really noise at all. For detectors of lesser capability internal noise forces approach to the grey area.

For sufficiently high light levels, particularly for high bandwidth systems, the dominant noise in a uniform level signal is the shot noise, typically approaching Gaussian statistics, and although the signal-to-noise ratio (SNR) improves, the actual noise increases with the square root of the power and the system bandwidth.

The optical power incident upon a detector is proportional to the square of the electric field, whereas photon noise typically depends on the square root of optical power or the electric field itself. In later circuitry, the electrical power is typically described by the product of the current generated by the incident light and the voltage across a load, becoming the fourth power of the electric field. A transimpedance amplifier, presenting a high input impedance to the detector with a low impedance to following devices, can provide wide-bandwidth high-gain linear amplification without seriously worsening the noise inherent in the signal.

After optical detection, electronic conditioning using pulse-height or time discriminators as photon counting circuits, or box-car analysers and lock-in amplifiers to average repetitive signals, can extract a signal seemingly buried in noise. Further improvements may be possible using auxiliary timing or other information that is independently known or can be derived from the signal. Techniques such as photon correlation can make good use of the photon arrival statistics, using information a light beam carries in the intervals between successive detected photons. For OADS, this permitted accurate speed measurements from correlated pairs of pulses as particles crossed two parallel sheets, whether the pulses were from individual or clusters of scattered photons, from particles of $<0.2\,\mu$m radius up to snowflakes.

Many real experiments fall between photon resolved and classical, and measurements are required where noise is a combination of statistical effects, whose relative importance depends on properties of the optics and detectors, as well as the phenomenon of interest. In this regime it is essential to consider all sources of noise, the relative importance of each, and their effect on the accuracy of the measurements. Simple calculations supplementing answers to carefully worded questions to device manufacturers improve confidence.

For single-photon detection, the signal is, in some sense, already digital, allowing various ingenious time dependent pulse processing techniques to examine the information encoded as variable intervals between one-bit level changes. Processing methods available here can be complex, as intimated by the information in the further reading list. Where detection yields an analogue signal, transformation to a digital form is preferable before further signal manipulation. An analogue-to-digital converter (ADC) produces a sequence of multi-bit words, immediately exploitable by digital processors. Transformation of an optical measurement to a usable digital signal usually demands low-noise linear amplification, DC offset control, and sufficient resolution for acceptable quantization noise, arising from the number of discrete levels, 256 with 8 bits, 4096 with 12 bits. Only for an extremely noise-free input or a large dynamic range is a 16-bit ADC currently justified. Where suitable averaging is possible, quantization noise can be reduced by adding known noise to provide random jitter between discrete levels, a technique used in RADAR. Most ADCs tend to be power hungry, and careful choice is necessary for applications where power is scarce, such as from batteries or in spacecraft. Noise specifications in two-dimensional arrays, such as CCDs, or addressable complementary metal oxide silicon (CMOS) arrays, can be less optimistic than at first apparent. Cooling can reduce the readout noise that dominates for CCDs, but excessive cooling can reduce sensitivity.

C5.5.7 Information retrieval

Two important questions are 'What is to be inferred from the measured optical signal?' and 'How is the inference to be drawn?'. Relating the value of the quantity to be measured to the detected light is not always obvious, as the subject and study of inverse problems attests.

Care is taken to optimize the signal in terms of intrinsic or additive noise, nonlinearity, over-ranging, etc. It is convenient below to distinguish signal and data processing, intended to manipulate and usually selectively reduce numerical data, from later information retrieval, weighted towards extracting the required quantities and their significance, to be presented in an intelligible form.

Signal and data processing

Signal and data processing may include precursor analogue manipulation, and the later rearrangement of a rich one-dimensional raw data stream into a more useful form. Whether from a single-point detector, or serially accessed pixelated sensor array, the data stream must typically be processed to emphasize the required measurement over irrelevant factors. Figure C5.4 is a symbolic diagram of the Weiner–Khintchine relationship linking the conservative Fourier transform (horizontal) with the selective discard of data (vertically down). Both routes from upper left to lower right are mathematically identical, but practically the route via correlation is often preferable because unwanted data are discarded sooner.

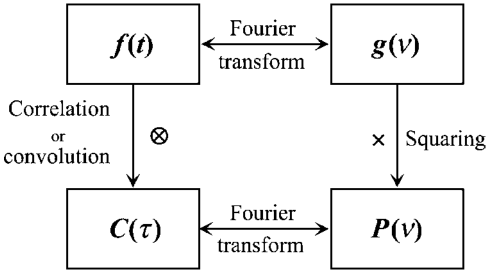


Figure C5.4. Weiner–Khintchine relationship.

Conservative transformation may make data more intelligible, whereas selective discard of the irrelevant may improve recovery of useful information. For the illustrative example of OADS, the analogue APD pulse from each particle transit was processed by automatic gain control and matched filtering. Successive sub-samples of each pulse were digitized and buffered by a first-in–first-out (FIFO) memory with conditional exclusion of digital samples not part of any useful pulse, obviating the need to process information-free noise between pulses. The FIFO was drained and processed to produce an amplitude and interpolated time for each probable particle transit, as a discriminated event. Infeasible sequences were removed by a conditional editor, again reducing processing load. A table of transit time pair probabilities, analogous to a correlogram, was accumulated in random access memory (RAM). Particles not crossing both sheets contribute a noise floor, while useful transits give a clear peak, whose centroid is parabolically interpolated to better than the 5 ns resolution available from the time tagging of pulses. This sequence of processes, combined with the parallel sheet geometry and its control, achieved the 1 part in 10^4 accuracy demanded, but not usually possible with optical anemometry. Accuracy was also aided by the absence of residual turbulence since the instrument was moved through stationary air.

Several processes must often be performed sequentially, and choices between real-time and off-line processing can be the subtle. If either the data input rate or the processing overhead vary significantly, then asynchronous interfacing with local buffering or even storage can be appropriate (section C5.6.6). OADS used the FIFO to tolerate wildly fluctuating input pulse rates, dual port RAM to accumulate correlograms, and conditional peak finding so that interrogation by the aircraft bus always accessed the most current measurement.

Estimating computing resources predicts bottlenecks to be alleviated by appropriate hardware. Machine architecture is tailored to provide sequential, pipe-lined, or parallel task execution, and assignment of the various synchronous or asynchronous tasks. However, the cost, risk, and suitability of available software can force hardware choices.

Many systems rapidly generate enormous numbers of data as one or more signal streams too rich to be recorded in real time. An example is a laser-based velocity sensor applied to a wind tunnel, an industrial flame, a ship's wake at sea, transonic flow, or inside an aero-engine. Correlation can compress the incoming event stream, selectively discarding only that which is less interesting, in this case, relative phase information. Where tens of thousands of correlograms may be acquired, often at great expense, a balance must be established between archiving all that could later be necessary and a reasonable storage expense, in time, money, or hardware resources. As a minimum, data as raw as possible, together with all system health and status monitors, should be stored and backed up as comprehensively as can be arranged.

Built and flown between 1989 and 1993, OADS acquired six incoming data streams each at $20\,\mathrm{Mbytes\,s^{-1}}$ to be processed by various dynamic and conditional discriminators to give three simultaneous channels, each of forty 32-bit correlograms per second with sub nanosecond delay time discrimination, for a flight of 2 h that cost several hundred thousand US$ (1993). This system used all of the techniques outlined above, and indeed many more. Data processing choices of how to display updates to the experimenter for real-time inspection and system control during research and development were also extensively different from those necessary for a final deployable system.

Information retrieval

Although distinguishing between data and information may seem merely philosophical, it can help to establish hardware requirements, and both usability and intelligibility of the reported numbers. While some systems are passive, perhaps even with data only available after the experiment, others may yield a real-time indication permitting user intervention, or possible redirection of the experiment.

By information retrieval is implied the recovery of the measurement in a format for archiving, with experiment provenance and monitors, and which is also immediately accessible to the observer. This implies that the presentation is meaningful to a human, important when real-time control or intervention may be necessary, or where proper experiment progress must be monitored or confirmed.

Software

For any instrument or system involving optoelectronic disciplines, an organizational methodology is necessary to make it all play together. Functional and structural design must be augmented by a software specification as early and as complete as possible. Without such discipline, a monster is created, a hostage to fortune, later to move out of control into cost overruns, chaos, and disillusionment.

Typically with optoelectronic instrumentation, signal and data processing and information retrieval will require software, as will the user interface, with its control, monitoring, and archiving needs. Software must be modular, testable, and appropriate for the planned hardware, although more usually nowadays the hardware is chosen on the basis of computing capabilities and resources demanded by software, which has become more expensive and less flexible than hardware.

C5.5.8 Interfaces

One critical area in any optoelectronic activity is the establishment and comprehensive definition of external and internal interfaces. Each interface must be compatible with both the elements that it is intended to connect. Isomorphic mapping of software and hardware modules (section C5.4 and figure C5.3) introduced the need for ideas expanded in table C5.8, without which testing and diagnostic

Table C5.8. Interface maps.

1.	*Operator interface* Command, control, access, reports, displays, archiving
2.	*Modularity* Block diagram and naming of functions, hardware, software structures
3.	*Information flow* Control, data, health and status monitoring, communication, formats, timing, protocols, archive, retrieval
4.	*Power and wiring* Input power availability, conditioning, connections Internal power supplies: average and peak voltages and currents, acceptable noise, cross-modulation limits Wires: type, gauge, insulation, colour Connectors: type, gender, shells, cable harnesses; runs, and makeup
5.	*Optical train* Power, wavelength, spectrum and bandwidth, polarization, acceptable aberrations, spatial and temporal coherence, noise, sources of stray light, ghosts
6.	*Physical structure* Mechanical: hardware systems and mounting structure, detail drawing index, assembly methods and sequence, available adjustments, alignment criteria Thermal: sources, sinks, pathways, effects, stability

analysis are made more difficult. An initial top-down design process is attractive, although it may often be necessary to accommodate bottom-up changes driven by available components, modules, subsystems or blocks of reusable software or code.

Accurate definition of all interfaces is imperative. Even, or perhaps especially, when other criteria may superficially appear to be more important in the establishment of system architecture, it is essential to resolve priorities to create consistent structure. Interface definitions start with a connectivity map of the proposed system, including its interaction with the outside world (e.g. figure C5.3), proceed via required functionality, ultimately to details of actual wires or numerical quantities to be exchanged, and the formats to which they are constrained.

The importance of these documents centres upon their practical use. Interface drawings are best as portable files in a format that can be read by everyone interested, and printed at least once as hard copy on a paper that will accept bold red marks, for peer or final review. The existence of interface diagrams as working documents, spread upon a conference table or a video-conference screen for examination and critical discussion, can greatly enhance the effectiveness of the final system, instrument, measurement, or scientific understanding obtained from its use.

Commitment of definitions of interfaces and connectivity to a formal document improves confidence that the system can perform its intended task, often suggesting simplifications and more logical ways of doing things, but more importantly defining modular boundaries. The specification of every module allows each part of the system to be tested independently of others, an ability whose importance cannot be stressed too highly. A second advantage is that critical components may be tested as soon as they become available, without waiting for everything else to arrive, not only for functionality, which is mostly reasonable for commercially purchased parts, but more importantly for fitness and agreement with expectations.

Within the larger context of the initially specified environment of the system, table C5.8 lists six general types of interface or definable structure. During design, each must be described down to a level sufficiently detailed to prevent inevitable later changes from rippling through the system with unanticipated effects. A hierarchy of mapped and documented interfaces reduces later problems, although some interference is to be expected between the tabulated items, expanded somewhat below.

Operator interface

The highest level, and in many cases the only one visible to the operator, includes the appearance of the system, the user friendliness of the available commands, the methods of reporting and displaying the required measurements, either in real time or recalled from archive, the confidence given by suitable values of health and status monitors, and a method of deciding what to do if something unexpected occurs. Despite the obviousness of this, many user interfaces could be better designed, for examples, videotape and disk players, digital cameras, and many computer operating systems. Typical optoelectronic instruments are complicated and are required to make difficult measurements. As far as possible the operator should be protected from this complexity, but nevertheless have immediate access to more detailed information should anything unexpected occur. Successful operation should need minimal or no intervention, presenting measurements succinctly and intelligibly, with a comforting status indication that everything is within its design range. If more control is desired, then it is available immediately on interrupt, which need not pause data acquisition, but only make more information or control available to the operator. This includes the ability to request more detailed monitoring of any function, to modify the existing control scripts, to recall archived data for comparison, and above all to have sufficient feedback to know what is going on either as a result of operator instruction or unanticipated factors. It can be expensive and embarrassing to have an experiment yield a record that is either unintelligible or has uncertain provenance.

Modularity

Most purchased optoelectronic parts or subsystems already have well defined interfaces. Even if they are not always the most convenient, it may be better to accommodate something that already exists. This can save time and reduce risk, but conversely much design effort may be consumed exploiting standard components and interfaces in a system being designed to satisfy unique needs.

The system overview document should be a connected block diagram where each block is labelled by function, hardware name, and software requirement, as an expansion of figure C5.3. On a yet more detailed view of this same diagram will be labelled the types of interconnectivity representing module control, information flow, health and status monitors, and supplied power. Even at the system level not all blocks need have all connections, and as the same discipline is applied to lower levels of modularity down to individual components, the connections per component will become fewer, simpler, and more specific. At the lowest levels of the functional hierarchy, many components will be passive. For example a lens, even though passive and appearing first in item 5 in table C5.8, should also be included in item 6, so that a later drawing may be created to specify the physical mounting structure and alignment requirements.

At each level the specifications of interfaces with the real world environment are again reviewed, understood, confirmed, and fully documented.

Information flow

Interface drawings for control, information flow, and health and status monitors are powerful tools. The thought devoted to creating these diagrams becomes the design work necessary to assure the complete and proper functionality of the final system. This is also a valuable method of checking internal consistency, and raising concerns that may even require customer choices. The optical part of this can specify everything about the radiation, the optical train, and optical components. For the electronic parts, except for standard communication interfaces, the actual wires that carry different types of information should be clearly identified to avoid confusion. Three separable overlays of connectivity are useful, (1) the command and control information passed to modules and components, (2) the health and status information returned from modules and components, and (3) signals, which also include the light, pertaining to the desired parametric measurements. In addition to their separate functions, the first two types allow closed loop control where this is desirable.

Active components have command and control capabilities with responses and properties monitored as often as necessary. Optical signals typically end at the detector, transforming to electrical signals, and thence to numerical signals for interpretation and archive, with inspection possible at various intermediate positions. Wherever possible, information exchange should be via standardized interfaces such as RS232, RS422, IEEE488, 1553, USB, IEEE1394, or many others. The five standardized communication layers of physical, data link, network, transport, and application should be understood, with special care necessary for real-time instruments.

Power and wiring

Power supplies deteriorate or fail in various ways, devolving from improper specification or performance all the way to poor mounting or inadequate cooling. A connector can be the wrong gender, sometimes the wrong connector, occasionally even with the wrong number of pins, and the actual voltages delivered in service may be anybody's guess. Careful specification of input power availability, instrument requirements, and bounding conditions can mitigate such mishaps. Auxiliary aspects must not be taken for granted—for example the occasional need for a trickle current to sustain settings—and are better discovered early to prevent the interconnect wiring from declining from a traceable colour coded minimal set into a rat's nest of post-assembly fixes.

Voltages and currents to various parts of the system may be quantified on the power interface diagram. Conditioning the output voltage and noise is the main function of a power supply, of which there will frequently be more than one. Each must accommodate maximum currents under all operational combinations and must be stable, with adequate freedom from noise, either intrinsic or introduced by switching or operation of some other part of the system. Where several power supplies are involved, they may interfere with each other to introduce mysterious noise properties whose spurious cross-modulation signals can masquerade as legitimate measured effects until properly diagnosed. Problems traceable to power supplies are not rare (section C5.6.7).

Specification of wire gauge, type, insulation, colour, and especially connectors should be consistent throughout the system, but will probably be compromised in purchased subassemblies. Heat-shrink insulation of each termination may be desirable, but does prevent electrical probing. Choosing the appropriate gauge, type, and insulation is obvious, and specifying an appropriately consistent colour coding is extremely useful. The author once had the opportunity to diagnose a cross-modulation problem within a customer's power supply where well over 230 wires supplied more than 54 different subassemblies from seven different types of allegedly independent power supply connected in parallel to the same aircraft bus. The compact wiring looms were tightly laced. Every wire was unlabelled, 16-gauge, Teflon coated—and white.

Specifying connectors at the top level of modularity, subsequently percolating down to every component for which electrical connectivity is required, as early as possible, prevents much pain later. Many connectors have long delivery times, and early definition of the type and gender of connectors, and the length and construction of cable harnesses, is valuable.

Optical train

The experiment in section C5.5.1 implicitly creates the interface diagram for all optical paths, and this evolves into formal documentation of power budget, wavelength spectrum and bandwidth, polarization, spatial and temporal coherence, noise, acceptable aberrations, sources of stray light, ghosts, and many other properties. In fact these may not all appear explicitly on the interface diagram, but are included here since a rethinking of these properties in the context of interfaces and compatibilities almost always identifies previously unconsidered items.

While poorly understood or excessive aberrations can seriously compromise performance, effort can be wasted reducing aberrations more than necessary. Monochromatic wavefront or ray aberrations can be generally represented by a polynomial expression of increasing order in aperture, field angle and azimuth and their products. The five third-order Seidel aberrations not correctable by a focal shift are primary spherical aberration, coma, astigmatism, field curvature, and distortion. Occasionally higher orders must be considered. Chromatic aberrations must also be considered in multi-wavelength systems. Unless the system demands special devices with exotic optimizations, good-quality components optimized for general use should be examined for suitability.

With many modern components, tolerancing is of more interest for performance prediction than correction, although as with ghost analyses it merits sufficient consideration to justify its subsequent dismissal. Where optical or physical path length is at a premium, substantial gains may be realized by folding the path. Although this may also improve rigidity and general robustness, there may be hidden penalties in design costs, performance, and flare.

Because in recent years, smaller, brighter, and more coherent optical sources have become commonplace, calculations based upon Gaussian beam behaviour are often necessary, and the

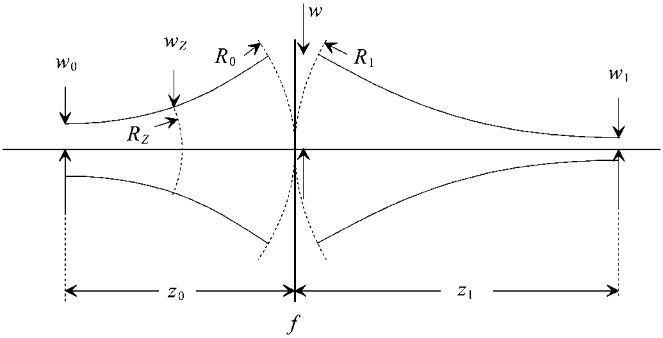


Figure C5.5. Gaussian beam parameters.

appropriate diagram and formulae are repeated here. Figure C5.5 and the inset box show the relationships of the beam radius and wavefront radius of curvature at a distance z from a Gaussian source of e^{-2} intensity radius w_0, with the asymptotic approximations in the far field. For a converging lens of focal length f at distance z_0 from the effective waist of the source, the remaining two equations give the aberration-free waist position and radius.

Beam radius w_z at distance z is

$$w_z = \sqrt{w_0^2\left[1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2\right]} \approx \frac{\lambda z}{\pi w_0}$$

Radius of curvature of wavefront R_z is

$$R_z = z\left[1 + \left(\frac{\pi w_0^2}{\lambda z}\right)^2\right] \approx z$$

Axial position of waist w_1

$$z_1 = f + \frac{(z_0 - f)f^2}{(z_0 - f)^2 + \left(\frac{\pi w_0^2}{\lambda}\right)^2}$$

Radius of waist w_1

$$w_1 = \frac{w_0 f}{\sqrt{(z_0 - f)^2 + \left(\frac{\pi w_0^2}{\lambda}\right)^2}}$$

Physical structure

Mechanical layout drawings are necessary for optical component mounting. Without them a draughtsman cannot detail parts, nor a machinist make them. Extending these drawings beyond the obvious also permits calculations of rigidity and, more usually for optoelectronics, ranges of necessary adjustments, confirmation of adjustment orthogonality, and locking techniques. Further refinements can establish a thermal interface map, including sources and sinks, and effects of conduction, convection, and radiation on the heat budget and temperature distribution. Thermal effects can compromise success, and show different behaviour in service from that noted during assembly, testing, calibration, and evaluation. In the absence of gravity, for example, convection is suppressed, and under different operating conditions very different thermal distributions can be expected, perhaps changing calibration or causing irreversible effects. Gradual drift from thermal cycling is common but not always suspected in its early stages.

Occasionally an optoelectronic system is so complex that a finite element analysis of the mechanical structure is a good investment, permitting mechanical and thermal loads to be simulated throughout any environmental range. Rarely do these programs interface smoothly with optical design programs, which themselves may involve more effort than the usefulness of the design refinement they can provide.

C5.5.9 Management and documentation

An individual scientist or engineer responsible for creating a simple optoelectronic device can occasionally track everything well enough to retain control using only a notebook and memory. However, this is rarely sufficient. The notebook is essential, and may evolve into a design document in electronic format. Four things are emphasized: that documentation should (1) be subject to configuration control from the first keystroke, (2) be backed up frequently and at more than one site, (3) be in a portable format that can be read by everyone who needs it, and (4) not become unreadable with time, either because it is stored on obsolete media, or because the generating software is no longer available. Few can now read $5\frac{1}{4}$ in floppy disks, and $3\frac{1}{2}$ in disks are rapidly going the same way. Even in comprehensive corporate networks, programs used to create and store information, particularly drawings, become obsolete surprisingly quickly, and translators disappear.

Documentation is key to success, whether by allowing design reviews to pre-empt problems, or by providing the necessary audit trail to diagnose and fix them. Table C5.9 suggests documents that support the success of the programme and the measurements.

The minimum programme plan is that which, in addition to its technical merit, will satisfy the customer about cost and schedule. Full-service project management software is usual for general programmes, but for optoelectronic activities and systems the match is rarely perfect, representing overkill in some areas and under-specification in others. A simple spreadsheet can be a valuable tool for keeping track of optoelectronic system development. For example, the first column can contain the nonorthogonal set of all components, activities, and concerns, followed by exemplary column headings such as specifications, source, supplier, address, universal resource locator (URL), phone, name of personal contact, cost, order date, promised delivery date, status, actual delivery date, tests to be performed, test results, assembly status, problems, comments, contingency plans, alternatives, mysteries, and more as design and implementation progress. More still will be added during building, testing and evaluation. This

Table C5.9. Documentation.

Programme plan	Cost and schedule plan, tracking of all activities
Configuration control	What documentation is current, and where to find it
Interface specifications	Modularity, information flow, monitoring, power
File backup	All documents immune to loss from single-point failure
File portability	All documents readable by all, now and in the future
Design document	A single source for all current data

spreadsheet is a convenient place to keep current information, which might otherwise be spread among many people, documents, mailrooms, desks, file drawers, memories, and wastebaskets.

Configuration control assures that information is current. Editing an obsolete file, repeating a task already completed, misplacing an unlabelled lens, or machining a part from a drawing that was changed in review but not sent to the machine shop, are ways to lose resources. Of particular concern is that purchased parts are often not exactly what the catalogue offered, and ripple-through effects can only be minimized by good configuration control. Everyone involved needs the most current information.

Comprehensive and powerful documents are wasted unless people have access to them, can read them, can understand them, and will act upon them. File portability and file backups are mostly handled by modern computer networks, but a poor choice of commercial software or unfamiliar modelling program can inhibit dissemination of, and access to, necessary information.

The primary responsibility for a design document is that of the technical champion who has assumed personal responsibility for satisfying the customer. Although in covering all levels from overview to the smallest detail, it can seem difficult to create and maintain, the power in comprehensive understanding of the final system justifies the investment. A convenient sequence of chapters might be (1) introduction and physical principles, (2) specifications, (3) system architecture, (4) interfaces, (5) performance modelling and prediction, (6) module function and physical descriptions (several chapters), (7) software, (8) testing and calibration, (9) power supplies, (10) table of all relevant and known numerical values, (11) method of contacting suppliers or other expert sources, (12) reasons for design choices, (13) log of evolutionary changes, (14) unresolved mysteries, and (15) several specific technical appendices. If this document is brought into existence at the beginning of the optoelectronic project and maintained merely by ordered accretion of all pertinent information, then it will prove invaluable. Abstracted sections can become the final report, comprehensive provenance for a scientific publication, the operator's handbook, the service manual, the diagnostic methodology, the platform for evolutionary improvements, and a delight to the customer.

C5.6 Components

Planning and design prepare the way for the choice and acquisition of components. In optoelectronics, the design process naturally iterates with what is available, and the loop between design and component choice is always tight. Components improve almost daily and innovations can sometimes solve problems considered challenging at the beginning. This is often magnified by a steep learning curve, where the uncertainty of initial unfamiliarity gives way to confidence as knowledge increases.

While section C5.5 is too brief to include all design concerns, this section addressing physical components must be even less comprehensive and current. Deterioration of relevance with time is exacerbated because quantities, qualities, nature, and even functional principles of the components available to the scientist or engineer practising optoelectronics evolve almost daily. To avoid becoming even more outdated than is unfortunately inevitable, rather than current device specifications, this section emphasizes questions and classes of properties that may perhaps continue to be relevant for the future.

Component choices are often guided by familiarity, cost, and/or common sense, any of which can be disastrous. Typically in any optoelectronics-based activity the cost of the components, however high it may appear, is largely irrelevant compared with the total programme cost. Buy the best and cry only once. In optoelectronics, and often optics in general, the common sense approach is often derailed by the intrinsically counter-intuitive nature of the subject. Even after many years of familiarity, what should be a naturally obvious improvement can often makes things worse. Conversely a less familiar component may seem attractive, if expensive. Rather than basing a choice on the manufacturer's recommendation, a better technique is to talk with someone who has used that component, ask the vendor for a demonstration, or borrow one and try it. Of course in some cases final choices may depend upon what we already have, or can get, or think we understand. Pragmatic examination of every aspect, reaffirming the validity of even well-formalized beliefs, carries high dividends. Check every assumption; check it again.

C5.6.1 Light sources

In a passive experiment, the phenomenon provides photons from which information is to be extracted. In active experiments using light to probe or intentionally stimulate a phenomenon of interest, the light source must have known, constant, and well-understood properties. The performance of the source may depend on properties of its final environment not necessarily suspected when the equipment is characterized in the laboratory. Where many types of light source are available, the simplest device satisfying the requirements is usually the best. Lasers have high intrinsic luminance, a desirable property for most applications, but where coherence is either not required or even undesirable, a super-luminescent diode may be better. Light emitting diodes (LEDs), cheap and available over a wide range of wavelengths, may be adequate. Increasing source power alone improves the signal less often than might be wished, and accidental modification of the probed phenomenon by a poor choice of light source is not impossible.

Type of source

The light source must have sufficient power and be stable under all likely conditions. Its properties must be matched to the system. Spatial coherence, implying the capability for good optical beam control, is usually desirable, whereas temporal coherence, essential for applications such as holography and interferometry, in others can be a serious disadvantage, giving rise to unwanted fringes or signal fluctuations unrelated to the measurement. Often not all the desirable properties may be available from a specific light source, and a trade-off may be necessary between, power, wavelength, coherence, polarization, modes, and their distribution and stabilities. The availability of a commercial product may even dominate the trade-off. Occasions where a thermal source is best are becoming less common with the increasing availability of many other ways to generate light, from innovative discharge lamps such as microwave excited sodium, through room temperature solid-state sources whose properties may be well specified over an increasingly wide range, to ingeniously contrived quantum-well emitters. Although much significant optical knowledge was gained with sunlight or wax candles, new principles of light production and photonic materials are currently driving the optical arts at an ever-increasing pace.

For OADS each of six thin sheets of light was imaged from a 1 W GaAlAs laser at 812 nm with 20 adjacent 1 μm by 3 μm emitters on the same die, with weak coupling to give constant output phase with a longitudinal intensity uniformity better than 20%.

Wavelength

Optimum wavelength is often an early consideration. The choice is rarely trivial since both the phenomenon and the optical system can force complex trade-offs. The spreadsheet technique introduced earlier to predict signals and noise levels can also be used to optimize wavelength, or indeed, any aspect of the system, by setting up a figure-of-merit involving all the parameters considered important and exploring the effects of various values, available components, and plausible changes.

Power and stability

Since optical power must be both sufficient and stable, it is highly desirable to either monitor the intensity (section C5.7.1) or better still use the monitor signal to stabilize the power with a suitable closed-loop controller. Despite manufacturers' claims, it is essential to characterize the source under the conditions of anticipated use, including ambient temperature and changes, power supply modulations, drifts and transients, mechanical vibrations and shocks, contamination, and ageing of construction materials. Although visual confirmation of spatial properties of the source, enhanced by special techniques if it is outside the visible spectrum, is often adequate, it is sometimes desirable to quantify the spatial distribution and transverse mode structure by a beam profiler or by image capture and analysis, requiring an additional two-dimensional sensor. If direction of polarization or longitudinal mode structure, particularly their effects on coherence, for examples, can change the measurement, then these too must be monitored and/or controlled.

Beam-pointing stability is particularly important in optical systems with high magnification or with strict directional requirements. A sufficient change of laser beam pointing direction can seriously compromise launching into a fibre, spatial uniformity of illumination, position of a measurement volume, or receiver alignment. Experimental quantification of effects increases confidence. Diode lasers typically must be temperature stabilized to retain power, wavelength, and pointing stability, with residual stabilization of polarization purity and orientation particularly important when launching into single-mode and/or polarization-maintaining fibres. The OADS lasers exhibited a negative wavelength dependence on temperature of about 0.3 nm $°C^{-1}$ and all six were kept near the centre of the 3.5 nm full-width half-maximum (FWHM) narrow-band filter, also temperature stabilized. Since each laser was slightly different the base temperature of each was individually controlled differently to give the same wavelength. A better technique would have been to use six different wavelengths and six separate filters to reduce signals from particles crossing other channels but the advantage was considered outweighed by the necessary increase of size and physical complexity.

C5.6.2 Optics

The term ‘optics’ conventionally describes the components whereby light is transmitted between source and detector via the phenomenon of interest. Wavelength is conserved in the most common cases, but inelastic processes such as fluorescence, or other nonlinear effects, change wavelength in complex ways. Optical components include lenses, beam-splitters, polarizers, fibre-optics, graded-index devices, coatings, stops, filters, attenuators, acousto-optic, Faraday or Kerr effect devices, windows, transparent adhesives, and many others. The choice from so many components should always be the minimum set of the simplest parts that will meet the specification. Avoid moving parts; avoid unusual or nonstandard parts; where possible, avoid parts. While the usual goal is to transmit as much light as possible that carries useful information, the reduction of flare can be as or more important, because excessive flare can mask desired effects and also consume bandwidth or processing resources. The optical properties of the system are crucial, for once information is lost it can never be recovered. Even if it is merely swamped recovery may be impracticable.

Lenses and other components

Competing criteria affect the choice of optical components. These include maximum transmission at the wavelength of interest, control of aberrations, implying not only reduction but also exploitation such as apodization, surface placement and curvature, and treatments designed to reduce unwanted reflections and scattering. However good a standard design, it must be evaluated for each application, and any sufficiently innovative requirement may demand special consideration. A particular example of this is to think that zoom lenses, whose compromises seem adequate for photography, might be applicable to optoelectronic instruments. Despite the tempting cost and apparent versatility the results are often disappointing and sometimes disastrous, because of the complexity and unsatisfactory compromises introduced to solve a problem not present in the system of interest.

Sometimes an unconventional component may promise, or even deliver, great rewards. However, excess attention to the improvement may hide shortcomings not apparent with a more conventional approach. All that may be said about lenses applies to flatware such as prisms, or a variety of more complex and unusual components such as diffractive or holographic optics, nonimaging concentrators, or other less common devices and techniques that may yet solve a specific problem effectively. Smaller components have better mechanical integrity and are less prone to distortion with mounting and conditions of use, but they can be more expensive and difficult to handle, and are more prone to impaired properties close to the edges. Even with well-made optics it is good to have the component at least 10–20% bigger than the active dimension. For Gaussian beams an operational aperture radius of at least twice the e^{-2} intensity radius of the beam is recommended to avoid intensity profile or wavefront structure changes. For conventional optical systems the presence of implicit apertures forced by component clear dimensions or edge imperfections may introduce effects in addition to, and different from, the intentionally designed stops. Vignetting, intensity roll-off with radius, can impair imaging systems. Compensation in software by radial scaling can reduce the usable dynamic range of the sensor and increase the effective noise.

Stray light

Reducing stray light is a black art. For mono-static systems, where common optics illuminate the phenomenon and receive the returned light, control of stray light is always a challenge. Finding all sources of flare demands a full design understanding of the optical geometry and component properties, which can still yield surprises. For bi-static systems the problem is alleviated by physical separation of illuminating and receiving optical trains, but even here high multiple-order reflections and scattering in any common parts, such as windows, can contribute unwanted scattered light comparable to the often extremely low signal from which information must be extracted.

Stray light from sources external to the instrument may often be reduced by baffles, optical labyrinths, stops, and/or filters. Spectral control using narrow-band or sharp-edged filters usually requires a collimated beam, even for the temperature-tuneable and no-longer fashionable Christiansen filter, where a cell contains transparent liquid and powder with different dispersion curves intersecting at the transmission wavelength. OADS used a temperature-tuned dielectric stack in a well collimated beam. If necessary, spatial and spectral methods may be augmented by temporal methods, such as a modulated source with phase-sensitive detection. This can vary from simple ac coupling with a tuned or lock-in amplifier, with phase- or frequency-locked loop capabilities, to driving the source with a pseudo-random code generator and correlating it with the detector output. Where flare does not have the same time characteristics as the signal, many orders of magnitude of rejection become possible, and in the last case range-gating is an additional bonus. Modern solid-state sources, for whose physical properties a nanosecond is a long time, especially permit these techniques and the associated improvement of SNR.

Internal flare is often more difficult to suppress, but is amenable to the usual techniques of high-quality components, properly chosen and applied coatings, cleanliness, stops, and the choice of surface shapes. A tunnel diagram can help understand systems with reflective components, where the optical system is unfolded about each reflection, and all possible stray light paths become straight lines. This is good for finding where best to put stops and baffles, and quantitatively augments the powerful technique of actually looking down the system from the detector end. This visual assessment of the system often finds sources of stray light, particularly in systems designed for visible wavelengths, where the dark-adapted human eye is almost as sensitive as typical detectors. Even looking through systems not designed for the visible can still warn of unexpected consequences.

Ray-tracing programs extend the tunnel diagram concept, allowing for surface curvatures distorting the tunnel. Almost all offer ghost analyses, and in many cases help to identify sources of stray light and allow its reduction by proper choice and placement of stops, for example. Optimization is less obvious than software designers might suggest and there is no substitute for thorough optical knowledge and experience. Using optical design programs without having a realistic expectation for what the result should look like can be dangerous. It is easy to mis-set a single variable, and produce a design that may be a poor optimization or even nonfunctional. One aspect of the art is that if something feels wrong, then it probably is. For the 150 mm diameter receiver common to all six channels of OADS, quite elaborate spatial and spectral restriction was essential to discriminate against direct sunlight.

Coatings

Coatings are applied to most optical components to minimize surface reflection, to confer spectral selectivity, or even to protect. Using multiple layers improves efficiency and anti-reflection effectiveness but can increase scatter. The trade-off depends upon a calculated figure-of-merit for the chosen system. Well-chosen coatings can increase the nongeometrical aspects of light gathering power, reduce ghosts and increase SNR, particularly in mono-static systems. Conversely, they can increase scatter and reduce SNR. Optical coatings are sometimes more fragile than the base material, preventing cleaning, but conversely can sometimes be chemically or physically protective. Less commonly considered coatings, for example, electrically heated indium tin oxide, decrease transmission but may prevent worse damage from condensation, as with a cool window in a condensing environment. Where surfaces are superpolished, coatings require extra quality control and handling to avoid worsening the expensively achieved improvement.

Windows

Preservation of optical access with clean, undamaged, birefringence-free windows is often mandatory, but is rarely trivial. Windows and optical access in general can be ruined by local melting, stress cracking, surface crazing, ablation, thermophoretic deposition, condensation damage, defects induced by the phenomenon, improper mounting, or even by physical accidents. Occasional manual or automatic cleaning may be essential, as may cooling to avoid deterioration, or heating to avoid condensation. Table C5.10 suggests some factors of interest when specifying optical access, but other techniques are sometimes necessary, for example a high-pressure gas curtain or liquid film may sometimes be superior to a solid material as a window to protect the instrument from the phenomenon or its environment. Sacrificial membranes can be continually replaced, such as those to prevent spalling damage in laser machining. Transmission losses arise both from bulk properties and from surface effects. Temperature dependence of adsorption and bulk absorption effects, stress birefringence or mechanical variability are determined

Table C5.10. Window properties.

Material	Transmissivity, suitability, availability, preparation, cost
Physical	Strength, hardness, damage resistance, distortion
Optical	Refractive index, homogeneity, aberrations, surface finish, birefringence
Coatings	Effectiveness, durability, scatter, cleaning capability
Mechanical	Thickness, width, shape, stress, strain, retention, sealing
Thermal	Expansion, conductivity, ablation, cooling, heating, embrittlement
Stability	Change of state, etching, chemical or biological attack, crazing, ageing

exclusively by the material and specified accordingly. Surface properties are more variable and less well controlled. Good material, polishing, coating, and the assurance of their constancy are the best ways to reduce attenuation, scattering, polarization effects, and changes with time. Superpolishing under liquid reduces scattering from surface defects, but later coating or improper handling may compromise this. A trade-off must be made between scatter, reflectance, and vulnerability to damage, as it is often impractical to achieve the best of all three simultaneously.

Transmission through one or more windows must not unacceptably change intensity, wavefront curvature, polarization, wavelength distribution, nor introduce scattering or detrimental reflections during operation. Note, for example, that in an uncollimated beam even a perfectly flat plate has aberrations.

If requirements are either mutually exclusive or demand an expensive material, then double windows with an intervening balancing medium sometimes offer a cheaper and more practical solution. For example, simultaneous exposure to extreme pressure and temperature is possible using a thin magnesium oxide window exposed to the heat, followed by a cooled gas at high pressure, and a subsequent thick float glass window. Edge-polishing and annealing reduce mechanical stress-raising weak points. Support and sealing are best distributed over one or more faces with a suitable material that retains its flexibility. Elastic materials are good for repeated cycling over temperatures tolerated by the material. Annealed copper is good for one time sealing, and reasonable temperature cycling. For higher temperatures and cycling ranges, internally pressurized stainless steel bellows work well, as do thermally compensated materials. For many applications an adhesive with slight residual elasticity is good, but a rigid ledge may be better to support mechanical load, provided that stress is acceptably distributed. Thermal as well as physical shock can break materials.

More expensive and exotic materials are not necessarily better. For example, specific conditions in a flame chamber test rig broke all its sapphire windows immediately. Short of spares, time and money I had a local spectacle lens maker cut replacement windows from a piece of float glass. Surprisingly, a simple set of these cheap windows survived through the several month test programme.

Optical fibres

Optical fibres are commonly used both to transmit information, and to sense various physical properties as a technique for measurement. Because these capabilities are mutually exclusive, optical fibres are less easy to optimize for any given application than might be expected. Most investment has been funded by telecommunication requirements in the 1.3–1.5 μm wavelength region. Although this knowledge and experience has wide utility in optoelectronics, the properties of interest, either required or accidentally exhibited, are not always well controlled nor characterized, since they are not necessarily those of most concern to the manufacturers. Great care is essential in examining the properties affecting the specific application. Table C5.11 lists questions best answered before commitment to a specific optical fibre.

Optical fibres essentially transmit intensity alone, but extremely complex effects can alter the relationship between input and output so that the fibres act like intricate optical systems, behaving both as integrated optics, and also capable of applications such as amplification until recently occupied purely by electronics. The existence of photonic band-gaps and special optical properties, derived from physical structure on the order of the wavelength, has a wide field of potential applications but is outside the present scope.

Transmission of information is possible in many ways, including simple intensity, polarization, wavelength or frequency division multiplexing, solitons, and individual photons. Two applications of fibres are common, as a sensor, requiring that the output shall depend uniquely on the sensed quantity, and as a channel, requiring that the output shall indicate only the required properties of the input. In the second case, filtering using prior knowledge can improve the signal. For example, a single-mode fibre of

Table C5.11. Optical fibre questions.

Is single- or multi-mode fibre necessary?
Which of many fibre types is optimal, or even adequate?
What materials are appropriate and available?
Are transmission losses acceptable? And will they remain so?
Does transmitted or external hard radiation cause deterioration?
Should the fibre maintain, select, or control polarization? How?
How is injection mode-matching to be established and maintained?
Is the fibre sheath opaque to ambient light? Does it matter?
Must cladding modes be suppressed? If so, how?
Can properties of the fibre introduce signal fluctuations?
Does high light intensity induce nonlinear effects or cause damage?
Will the fibre be properly sensitive to the measured properties?
Will the fibre be properly insensitive to ambient properties?
What adhesives are usable or necessary for construction?
How shall terminations be arranged, aligned, and controlled?
How do vibration, bending stress, temperature, and ageing affect performance?

sufficient length to attenuate cladding modes can operate as an excellent spatial filter. Since a typical single-mode fibre remains single mode over a factor of two in wavelength, injecting the output from a superluminescent diode gives an excellent temporally incoherent point source, an important device until recently available only as an inferior approximation. Replacing the source with a laser permits retention of temporal coherence and again gives a versatile point source with negligible mass to align accurately. In the latter case, however, earlier comments must be augmented, and the fibre may have to be terminated with a wedged face to prevent coherent reflections from generating instability in the laser cavity or elsewhere, since the fibre can act as a sensitive Fabry–Perot interferometer. An inclined face changes fibre emission geometry from circular to slightly elliptical, with a pointing direction no longer parallel to the fibre axis. Dome polishing the fibre end can also mitigate detrimental effects of back-reflections, and facilitates compression coupling.

Launching into a single-mode optical fibre requires matching convergence angle or numerical aperture and waist size to preserve the Lagrange invariant, the product of aperture, field angle and refractive index, similar to the Abbé sine condition or the brightness theorem, which determines the highest possible freedom from loss of light gathering power. It is rarely possible to meet this condition totally, even with accurate and stable alignment, and birefringence and polarization properties of fibre and other components may worsen it further. Once light is launched into an optical fibre, it is not desirable to let it out and then try to re-launch it—the power loss is always worse than expected. Contamination of end-faces exposed to an open atmosphere can massively reduce the amount and quality of transmitted power, or even destroy the component. Emergent power density near the fibre face for one transmitted watt of green light can be $10^{11}\,\mathrm{W\,m^{-2}}$, eight orders of magnitude greater than sunlight. This may be sufficient to attract dust by thermophoresis, causing enough local heating to craze or ablate the fibre material. Such problems are easily recognized. Output damage is identified by the appearance of Airy-type rings or other intensity structure within the ideally Gaussian output beam in the far field. Input damage causes only loss of power. Optical fibres are typically fragile, requiring one or more tough concentric sheaths outside the cladding for protection. Their use also demands high-precision mechanical designs and mounts, typically to sub-micron and micro-radian tolerances. Even standard connectors for single-mode optical fibres have complex and exacting requirements, and transmission efficiency can fall with each reconnection.

Transmitting high optical power, especially through single-mode fibres, requires careful specification of fibre material, such as a germanium-free core that will not form colour centres, and special optical adhesives that do not denature. Transmission may reduce with exposure to high transmitted power or cosmic or other radiation, micro-cracking with vibration or under sustained strain from too small a bend radius, or inadequate physical support. Unless intentionally exploited, sensitivity to external effects is detrimental. Microphonics may sometimes be reduced by special care with alignment, and sometimes not, depending on fibre and environment, either of which may be insufficiently well specified. The only way to assure adequate performance is by testing a sample of the specific optical fibre from the batch available.

The conjugacy of end faces of coherently bundled fibres allows imaging of relatively inaccessible situations from body cavities to turbine blades. However, beware of the potentially ambiguous meaning of ‘coherent’ as applied to fibres. A coherent bundle is a loom of fibres whose ends map spatially so that an image may be transmitted, whereas a coherent fibre typically conserves spatial and/or temporal coherence of the light. Optical fibres based upon coherent constraint of the propagating modes, by step or graded refractive index control in regions of a few microns diameter, are distinguished from incoherent optical light guides, typically with step index confinement within a few tens of microns up to millimetres in solid material or liquid-filled tubes.

A special case is the GRaded INdex (GRIN) lens exploiting the radial control of refractive index to perform the operation of miniature lenses, using the techniques of optical fibres. Control of refractive index gradient in fibres may also inhibit dispersion and increase bandwidth. In a step-index single-mode fibre an evanescent wave extends into the cladding material. Ambient conditions may be sensed by their effect on this evanescent wave, and hence on the propagated intensity or polarization. This same sensitivity manifests as detrimental effects for pure transmission, where bending radius, temperature changes, stress micro-cracking, vestigial inhomogeneity from irregular sleeking near the core–cladding interface, and other manufacturing imperfections can give rise to a range of curious effects only fully appreciated by exhaustive empirical testing. Polarization rotation, temporal fluctuations over a wide range of frequencies down to drifts over periods of weeks, and cladding mode coupling are all possible. Used with GRIN lenses that exhibit residual birefringence, the fibre can behave as a temperature dependent wave-plate. Sometimes fluctuating light leaks are visible where the sheath is not opaque. Even with polarization-maintaining fibre, using bow-tie stress birefringence or elliptical cores, a slight error in launch orientation can yield large and unexpected sensitivity to fibre environment, and not necessarily in a predictable way. Polarization preservation where one of the two orthogonal axes is attenuated gives a purer polarization output, but occasionally at the expense of intensity instability. As fibre manufacturing control and methods of characterization improve, many of these effects may be reduced, or at least their sources better understood and avoided. Meanwhile specification for a given application must allow for experience specific to the individual fibre type, manufacturer, and indeed the production batch. Six 1.2 mm diameter clad quartz fibres were used in OADS as light pipes following a single chrome etched field stop to direct each of the six sheet images to its own APD.

C5.6.3 Physical mounting

Physical mounting of components must maintain correct position without stress or distortion. To avoid stress concentration, support must spread the load either by flexible media or sprung clamping. The hole, slot and plane mount, with gravity or light spring loading, provides kinematic location but cannot be locked without over-constraint. It is thus useful for exact relocation, for example of a hologram for reconstruction, or a component that must be removed and then replaced exactly, but is not robust and is typically confined to the vibration-isolated stable table of a temperature controlled laboratory. Even in that environment, where optics are mounted on posts, the posts should be as short as possible with the largest possible diameter. Where a mount is fixed to a supporting metering structure with screws, the component should be dished to provide the largest footprint, and controlled locking torque specified to constrain distortion. While the metering structure may be as simple as a flat table with tapped holes or magnetic clamps for greater versatility, for specialized instruments it may be a complex casting, machined frame, or composite assembly of diverse elements. The retention of its physical properties, especially dimension and rigidity, is important. In OADS the necessary beam pointing stability of $<5\,\mu$rad was achieved by a metering structure of low-expansion stainless steel, with weight and rigidity optimized by finite element analysis of the complete structure, isothermalized by ethylene glycol circulating in labyrinthine manifolds. Each laser was precisely located in one of the six circularly symmetric transmitter lens assemblies with five lockable adjustments, two tilt, two transverse, and one focus. Each completed transmitter assembly was directed to the correct point in space by rotation in a hollow spherical bearing near the output nodal plane, using one rotation and two more tilt adjustments. Alignments were made in a set sequence to meet specified criteria. The adjustments were locked by screws to a specified torque and all critical sliding or pressure points encapsulated in rigid adhesive. Most applications require neither this precision nor its associated complexity.

Whether in benign or more demanding environments, there is a trade-off between the rigidity of accurate placement and available adjustment to accommodate changing conditions. Active control is complex, to be avoided if possible, but sometimes essential. Large or fragile components may be held in a sine-wave mount, a perforated annulus of elastomer. To seal and/or distribute stress, O-ring seals or well-chosen flexible adhesives often suffice. Metal clamps are hazardous unless appropriately sprung. Where no movement of the component can be tolerated, rigid sealing may exploit materials with graduated properties to distribute the stress.

Ideally, optical mounts must be robust and rigid, without affecting the optical properties. For materials with optical activity or birefringence induced by stress, this can be confirmed by viewing the component between crossed polarizers, often an instructive test, with surprising results if applied, for example, to spectacle lenses. For high-quality components even small mounting stress can introduce aberrations, and rigid retention for high performance competes with more flexible location for field survival.

Adhesives are common in optical packaging, from component retention to optical function. Several popular transparent epoxy-based materials can be cured by ultraviolet exposure; others require two components, a catalyst, heating or time to reach the required state. When used for retention, the final mechanical properties dominate, but when adhesives are in the optical path more careful choice is advisable. Transparency is usually necessary, and this may be compromised in use by ageing, nonoptical radiation, stress cracking, change of polymer properties, crystallization, or high light level. The last is especially relevant where, for example, an illuminator has a GRIN lens glued to a single-mode fibre.

C5.6.4 Detectors

Ideally, any square-law detector yields an electrical signal corresponding to incident intensity, manifest in the semi-classical approximation as successive quantizations, which may not be resolvable. For sufficiently low light levels, an appropriate detector reports a given fraction of individual photon arrivals, missing some but not changing the arrival statistics. As the light level increases, this becomes a semi-continuous signal whose shot noise increases with the intensity, approximately as the square root of the equivalent number of photon arrivals, improving SNR. An audible analogy of this might be hail on the metal roof of a noisy workshop. This is true for a single-point detector such as a PMT or APD or for one pixel of a CCD, CMOS or other array. Criteria for optimizing detector type assume different relative importance depending on the application. For photon correlation, the ideal is a single-point detector free from dead time, after-pulsing, pulse pile-up, internal correlations, and with a sensitivity and quantum efficiency as high as practicable, to produce single resolvable pulses from every field quantization, with no triggering events from any other source. This may be approached quite closely for high-energy ultraviolet quanta with a rubidium telluride photocathode, as with OADS pressure and temperature sensing. For red photons both energy and cathode sensitivity fall, allowing more false triggers and impairing the correspondence of pulse statistics with those of the optical field. Single-photon performance is approached with reverse biased silicon APDs in the Geiger mode. To obtain highly accurate measurements of high intensity at low bandwidth, the criteria may shift from shot noise intrinsic to the signal to other sources such as Johnson noise from the load resistor or pre-amplifier, and keeping the noise low moves from optical to electronics design expertise. Cooling the detector and its electronics below a certain temperature often helps with noise, more so where the detector relies upon absorbed energy as with a thermopile or bolometer.

For single-channel detectors, PMTs are fast and can have huge sensitive areas, with the penalty of being physically large and requiring high voltages for the accelerating grids. However they offer amplification of many orders of magnitude up to thousands of amperes per watt, without adding excess noise to the original quantum realization variance. Wavelength sensitivity typically falls from mediocre (~25%) in the blue and green to rather poor (~1–4%) in the red, although recent implementations claim significant improvement. APDs have much better quantum efficiency (~90%) in the red, and are physically very small. Although their noise figure can be as low as 10^{-14} W $\mathrm{Hz}^{-1/2}$, when stably cooled to around 0°C, they also require a moderately high (~200+ V) and stable voltage to achieve a gain from a few to a few tens of amperes per watt. The sensitive area is often inconveniently small (<1 mm^2) with a longer dead time than PMTs. Lower-performance bulk detectors of various materials are often cheaper, and may be adequate for less demanding applications. Operating in either photovoltaic or photoconductive modes without intrinsic gain, their noise typically increases with surface area and temperature. Most detectors are fairly robust to optical overload, but take varying times to recover. For example, a PMT will typically recover its sensitivity lost from transient optical overload as soon as the dynode decoupling capacitor chain recharges, but to recover its noise characteristics may take several days in darkness with full voltage applied. APDs flip into an avalanche overload state if the optical input exceeds a level not much larger than the average, but are not damaged if the current is externally limited. Recovery after the removal of excess light can be quite rapid, but must sometimes be encouraged by the temporary complete removal of the reverse bias voltage. Thus both PMT and APD detectors can be protected by self-limiting mechanisms, and are not necessarily permanently damaged until the optical input melts the sensing or amplifying material. This is also true for conventional bulk material sensors, although, without limiting the electrical current, thermal runaway can cause permanent damage.

Array detectors operate on similar photoelectric mechanisms but store charge locally for an equivalent exposure time. CCDs read the charge out serially via a bucket-brigade, which is the major source of noise at low light level. Individual pixels rarely have uniform sensitivity and a calibration table is desirable for quantitative work. Ageing and radiation damage also change the characteristics of individual pixels, demanding calibration and update of lookup tables. Types of array detector are now available with different characteristics such as pixel size and shape, dead space between pixels (fill-factor), well depth (how much charge can be stored at an individual pixel site without nonlinear effects), readout speed and method, and even manufacturing technology. For example, CMOS requires less power and can be constructed so that each pixel may be separately addressed, clocked, and amplified. This conditional interrogation can greatly increase the dynamic range of the array, because those pixels that are more brightly illuminated may be read out more rapidly, accumulating the total from each pixel in external memory. Pixels in the darker areas are merely read less frequently or even at the end of the image acquisition time. In this mode, low pixel crosstalk is essential. While excessive cooling can cause a CCD to stop functioning, the effect on a CMOS detector is desirably to reduce the noise. Typically the output is read to an ADC, either on or close to the detector chip. While the digitization noise might appear to be no more than the inverse of the number of bits, real devices rarely perform this well and a bit or two of additional noise is common. Claims for the relative merits of each type of device change almost daily and for any application it is important to understand how the manufacturer's claims relate to the proposed experiment.

The art here is to choose a detector that seems appropriate and become as familiar with it as possible, including how best to drive and control it, how to compensate for its shortcomings, and what its limitations dictate for the accuracy of the experiment and the quantitative validity of the conclusions.

C5.6.5 Miscellaneous observations

Light gathering power can easily fall below the upper limit set by the Lagrange invariant unless the numerical aperture is matched everywhere. This occurs readily when coupling single- to multi-mode fibres, or at the entrance to a detector, under the impression that an alignment criterion may be relaxed. Typically the receiving optics, collecting light scattered from the probed phenomenon, is merely a photon-bucket, but this does not necessarily mean that aberrations can be tolerated; for example, the proper performance of a field stop may rely upon its high-definition image at some other location. Aperture stops may be similarly critical. The performance of dielectric stack narrow-band filters is typically compromised in a poorly collimated beam.

A given specification may be met and implemented by many different candidate components, devices and the phenomena upon which they rely. New techniques and innovative devices are constantly becoming available. Naturally occurring materials are now being augmented by artificial structures, including sequences of differently doped layers, physically and/or chemically controlled photonic band-gaps, and many other ways of spatially or temporally manipulating light from classical to quantum domains. New manufacturing methods, extending beyond chemical vapour deposition (CVD) and molecular beam epitaxy (MBE), are being introduced. Although fashion, implying ready availability and perhaps low cost, favours the new, many older ideas also have merit. Having at least heard of as many obscure or archaic devices and techniques as possible can be very useful, where old ideas may find renewed application because of other perhaps unrelated advances. For example, etching diffraction gratings from laser created fringes exceeds the capability of the once necessary complex and precise mechanical ruling engines. A broad knowledge of natural history is beneficial, since many subtle optical and electronic ideas have been common in nature for millions of years. For example, the Bayer mask used to achieve colour sensitivity in single-component CCD arrays without serious resolution penalty is similar to the implementation of colour vision in the pigeon by colour masking of retinal cones. The eye of Anableps has bifocal lenses to give clear images above and below the water surfaces it favours. Contraction of a cat's pupil to a slit permits a larger dynamic intensity range than a circle would, without impairing the motion sensitivity of an already fovea-deprived organ. However, other than as examples, such details are beyond the present discussion of arts and ideas to enhance success.

C5.6.6 Electronics

The transition from detected photons to sensible information is mostly performed by electronic devices, starting with amplification, via digitization, signal processing, data reduction, to retrieval of information for archive and/or presentation as a human-readable display. The hardware implementation of ideas introduced in section C5.5.7 has expanded to capabilities unimaginable a few years ago. Since this trend seems likely to continue, a discussion of devices is less useful than brief comments about general techniques, although even those become more comprehensive with each passing publication. An important trend is that optical techniques and devices are increasingly able to perform tasks once exclusive to electronics, from the extreme example of optical fibre replacing copper wire, to the integration of optical components with architectures formerly used only for electronic integrated circuits. The discipline of photonics is paralleling that of electronics, exploiting nanostructures with capabilities as yet barely imagined.

Once the detector has responded in whatever terms are available, the transition to a digital format should follow as soon as possible, usually via a pulse discriminator or ADC. Although the same bandwidth product of data throughput may be represented as a rapidly varying 1-bit signal or as a less frequently updated multi-bit word, available hardware often suggests an optimal operation between these two extremes.

For real time systems where the input is synchronized with an independent activity, bottlenecks impose undesirable limits, and techniques such as dedicated hardware with a real-time operating system may be necessary, perhaps demanding low-level coding and testing, even with integer arithmetic and/or bit-sliced architecture. However, modern chips currently appear so fast that even a more versatile engine may be adequate and current advances in hardware exceed those in software. However, that has not always been so and the situation may change again, and always the choices made to achieve system function are susceptible to critical review.

Where delayed throughput is allowed, asynchronization can be achieved by a FIFO buffer, through single- or dual-port RAM, to hard disk files in a suitable format for later access, depending upon the required buffering and accumulation rates. More than one asynchronous interface may be needed to retain clean modularity in some complex systems, and to complete specified tasks within the time available. Hardware choices vary widely, from application specific or very high-speed integrated circuits (ASIC or VHSIC), through digital signal processors (DSPs), via field programmable gate arrays (FPGAs), to generalized serial microprocessors. Each has its pros and cons. ASICs and VHSICs are fast, expensive and inflexible. DSPs are fast, versatile, and typically externally pre-programmed. FPGAs can be slower, but are dynamically, even conditionally, re-configurable. Microprocessors are completely versatile, but with the most common operating systems cannot process signals in real time because of intrinsic and uncontrolled task timing. Real-time operating systems mitigate this, but application software is less readily available. At each stage of the processing one wishes to have fewer data to handle, but can perhaps tolerate more complex processing algorithms. Hardware and software tools evolve rapidly. They become faster, more powerful, and more versatile, but also more complex, more difficult to use, and often have reduced backward compatibility.

Processing algorithms and hardware must be optimized as a system, retaining modularity and testability. Analysis is based on information intrinsic to the signal, but can be augmented by supplementary data to extend the limits of what information can be extracted, and the boundaries of attainable resolution and accuracy. Exploitation of independent auxiliary data, or any of a wide range of mathematical and/or statistical techniques, such as maximum entropy, analytic continuation, and super-resolution, may improve results. Understanding the output is not always trivial and iteration between the design concepts in section C5.5.7, and the hardware choices here, may be necessary and can be challenging.

For OADS, the 200 MHz raw signal was dynamically scaled and filtered before the ADC, followed by discrimination and conditional editing implemented in discrete components prior to loading into the FIFO, subsequently accessed by a TMS320C25 DSP, using scaled integer arithmetic with a real-time operating system. Its program was downloaded on initialization, permitting easy upgrades during development, and later evolution of function. For example, velocity, pressure and temperature were reported during twenty-four 1 to 2 h F-16B flights in 1990. Velocity and particle properties were reported during later research flights on F-104 and SR-71 aircraft, with final emphasis changed towards particle size and concentration measurements on final DC-8 flights.

C5.6.7 Power supplies

Without sufficient attention, power supplies can be more risky. Not only can they fail, but they may also induce into the signal instabilities, correlated fluctuations, noise in one or more frequency bands, and inconstancy with input voltage, temperature, load, and/or ambient conditions. Although these are mitigated by careful design, specifications must be confirmed by testing, calibration, characterization and monitoring of any artefacts in the light source or signal. For thermal sources, inertia prevents significant fluctuations faster than a few tens of kilohertz, but solid-state sources can have sub-nanosecond fluctuations. Accordingly great care must be given to the power supplies driving the light sources, detectors, and sensitive electronics. Although nontrivial, feedback stabilization should always be considered and may often be necessary.

C5.7 Testing and calibration

Much can be learned from just looking at the system and phenomenon, and even incidental observations may indicate unexpected behaviour, whose rectification builds confidence in subsequent measurements and the validity of conclusions. Components should be rigorously tested as soon as available. Curious effects are often noticed near the limit of awareness, and although they may be neither repeatable nor completely understood, their dismissal as insignificant may invite later catastrophe. Table C5.12 augments table C5.6 suggesting further questions whose proper answers may improve confidence.

No amount of testing can ever prove that software is infallible. The likelihood of unknown error is high. Good designs allow upload of new code modules to correct anomalies found in testing, permitting analysis and rectification before damage occurs. Power supplies must not only be tested as individual units in isolation, but also in their final operational configuration. It is well to impose test conditions outside those expected, and for longer periods, logging all relevant parameters with higher accuracy than seems necessary. It is important to examine these test results in various graphical formats and actually think about what trends might mean. Surprisingly many failures have origins in data measurable, or even measured, long before.

Tests should simulate operation as closely as possible. In OADS as a subset of system testing, the lasers could be independently modulated and the response of each APD to its own and other channels examined for various targets. Complete test documentation allows later insights about why the results are not identical with what was expected—they never are—and can allow compensation or correction to improve the measurements or performance.

Calibration is based on knowledge gained by testing. Quantitative characterization may permit adjustment of output values to reflect their input origins during operation, or to correct measurements later. The two aspects of calibration are the quantitative aptitude of the apparatus for its design purpose, and the traceability of the recovered information to required properties of the examined phenomenon.

Table C5.12. Testing and calibration.

Initial function	Does everything work as planned? Does it meet design expectations?
Anomalies	What do unexpected observations mean?
Deterioration	In a simulated environment, does it get worse at an acceptable rate?
Contingencies	How is less-than-ideal operation to be accommodated? How is the unexpected to be handled?

For OADS, the only important physical calibration was the separation and parallelism of the sheet pairs, performed by exposing a dimensionally stable sensitive film to the illuminated pattern at several axial stations and using a travelling microscope to measure the geometry.

C5.7.1 General health and status monitoring

Extensive monitoring is desirable. Power on system (or self) test (POST), and routines to exercise the equipment as it approaches operational status, are conventionally provided, but should be augmented by health and status monitors for all quantities that could affect performance or the values of retrieved measurements. Monitored values should be recorded as a time-logged data stream at whatever rate is appropriate for the parameter concerned. Table C5.13 lists examples of usefully monitored quantities with supplementary notes following.

For temperature, pressure, and thermal controls, a readout rate of 1 Hz is mostly adequate. Power supplies require either a more rapid readout, or a separate assessment of rms fluctuation, noise, crest factor, or other warning of unexpected operational noise or transient spikes. If the illumination intensity can fluctuate on a timescale shorter than the inverse monitoring rate, then additional properties should be assessed, as with power supplies, above. Most serious may be fluctuations and drifts on a timescale similar to that of the detector or camera exposure or readout, but not necessarily in phase. This can lead to quite subtle effects in later processing or, for two-dimensional sensing, picture-to-picture variations, and moiré or other patterns in single pictures. If such obvious effects persist despite design approaches intended to remove them, then the bad data may be rejected later only at the cost of efficiency, if they may be rejected at all. However, small attendant changes in SNR in the detector signal or picture are not so easily discovered or compensated. Indeed, these effects may be almost impossible to compensate, so it is important during design and testing to eliminate or at least reduce problems associated with the interactive effects of power supplies and the devices they drive. For example, calibration may change if only parts of a multifunction apparatus are currently activated, and this must be reported in the health and status archive attached to the provenance file for the specific experiment. Monitoring helps support claims for the validity of measurements.

Many internal systems merit stabilization by negative feedback, with the loop signal monitored. With modern computing systems it is relatively cheap and easy to characterize such behaviour. Experimental data may be only poorly understood without at least a time record of rms and peak-to-peak fluctuations, together with the amplitude and phase, or even power spectrum of the light source noise and drift.

Table C5.13. Candidate health and status monitors.

Power supply input and output voltages and currents
Illumination level and stability of the light source(s)
A sampling of temperatures throughout the apparatus
Internal pressure, humidity, and other vapour pressures
Electrical and thermal control voltages and currents
Fluid pressure, temperatures, flow rates, leaks
Loop signal in negative feedback stabilization wherever used
Independent views of the phenomenon and instrument (CCD cameras)

C5.8 Hostile environments

Optoelectronics can be challenging enough even in benign conditions. Hostile environments that threaten success are of two kinds, hostile to the instrumentation, and hostile to the observer. These include all aspects of the measuring process during operation, and the programme that brings the measuring device to the current situation.

In the first class, conditions such as extremes of temperature, pressure, radiation, acceleration, vibration, shock, chemical attack, electromagnetic fields, noise, power source fluctuations, cosmic radiation and perhaps other specific conditions lead to potentially less than satisfactory results. Even if the system is designed to undergo graceful degradation, successive levels of deterioration may include (1) loss of calibration leading to poor or untrustworthy measurements, (2) loss of function, preventing any measurements, (3) destruction of the instrument, or finally (4) collateral damage, implying that the instrument is not only destroyed but also causes associated damage that may be much greater. The loss of a $1.5b Mars mission because the same type of vulnerable component was used in all three of the triply redundant timing circuits was a sad example of instantaneous and unexpected failure.

Three aspects of environments hostile to the designer of the system are (1) not enough money, (2) not enough time, and (3) not having a champion. Although the first two are common and obvious, the last is the most serious. To make anything happen takes someone dedicated to its successful completion. During instrument deployment the observer may also be exposed to hostile environments, such as extremes of temperature, pressure, vibration, acceleration, radiation, noise, antagonism, unreasonable expectations, or stress.

Four classes of hostility merit consideration during the design of the instrument or methods for its deployment and use. These are where (1) conditions in the phenomenon are extreme, (2) access may be constrained, (3) the instrument is exposed to an extreme environment, and (4) the observer is under difficult or stressful conditions. Table C5.14 introduces ideal properties whose consideration may improve success.

Obvious though they may appear, warnings in Table C5.15 apply in benign as well as hostile and environments and may apply to the instrument or the operator, or both.

Outdoor or field installations are often exposed to diurnal temperature cycling or steep temperature gradients, whether in use or not. Direct or reflected sunlight, rainwater, and unsuspected fixed or pulsed magnetic fields can introduce problems. Exposure to moisture or high humidity can age optical surfaces and dielectric coatings, also possibly causing fungus build-up or other biologically driven damage surprisingly rapidly. Any one of these is easily averted or accommodated: avoiding them all, and also the unanticipated ones, can be challenging.

In summary, optoelectronic measurements in hostile environments are difficult. Analysis and thoughtful design mitigate risk. Planning graceful degradation, anticipating the unknown unknowns, talking honestly and often with the customer, and having plenty of money, time, support, and good luck are all beneficial.

Table C5.14. Ideal properties.

Simple, low technology	Easy to understand
Rigid, robust, small	Resistant to damage
Modular, reliable, cheap	Easy to diagnose and fix
Automated by default	Reduces operator mistakes
Low power demand	Stays cool, widely applicable
Calibrated, verified	Constant and trustworthy
Tested, characterized	Known and understood

Table C5.15. Warnings.

Chemical attack ruins	Noise reduces capability
Gravity changes disturb	Hard radiation damages
Vibration changes things	Power supplies fail
E-M radiation gives errors	Windows get dirty
Stress reduces performance	Software crashes
Temperature drifts misalign	Pressure distorts
Changes affect things unexpectedly	Fluids leak

C5.9 Checklists

Table C5.16 lists some component properties deserving examination, followed by table C5.17 stressing environmental aspects.

C5.9.1 Analysing unwanted effects

Any part of a system or its environmental exposure may have unanticipated effects on any other part, and to minimize later surprises a useful technique is to create a spreadsheet whose first column contains all possible affective properties and whose first row contains all aspects that could suffer effects. Examples of both properties and components are given in table C5.18. In the body of the spreadsheet, a manual entry is made into each and every cell to verify that any mutual effects between row and column heading have been considered. Effects here include any possible interaction or sensitivity, which may then be quantified, or less tangible warnings, such as excessive complexity. The exercise may trigger a memory of a similar or analogous situation with good or bad former outcome. The cell entry creates an audit trail later to become a powerful diagnostic tool. If you think this is a tedious and unnecessary exercise, try fixing things later. Most entries will indicate that there is no cross effect. Other possible entries are that the effect has been considered and is acceptable or that the effect has been analysed and remedial action taken, as indicated by a traceable reference. Perhaps most seriously, and a major justification for this exercise, is that an effect is found to be likely but unknown, requiring further effort directed either towards its understanding and rectification, or its quantitative inclusion in the risk analysis.

Table C5.18 is not intended to be exhaustive, but merely contains an exemplary set of possible subject items to be reduced, refined and/or augmented according to the specific optoelectronic system.

As an example, a reduction in ambient pressure once seriously compromised the behaviour of an argon ion laser. With the reduced refractive index of the lower-pressure air, the emission angle from the intra-cavity. Brewster windows changed sufficiently to detune the resonator cavity. Unexpectedly, retuning did not restore the power since the new direction in the cavity passed through a slightly thicker section of the wavelength selection prism, also intra-cavity. Even though the increase in absorptive attenuation was slight, the many passes within the resonator substantially reduced the cavity gain. The first lesson is that reduced ambient pressure changed the pointing direction and power in a way that was not immediately recoverable by realigning the cavity—the prism had to be physically displaced. The second lesson is that since such ion lasers are no longer in common use, the first lesson is largely irrelevant. The real lesson therefore is that a situation must be examined in greater depth and with more imagination than might at first seem necessary.

In a working career thousands of such examples appear, but using the cross-reference effects matrix postulated above can improve the probability of creating a working design with minimal evolutionary iterations.

Table C5.16. Component considerations.

Optical radiation source	
Type	Thermal, laser, diode, flash, arc, discharge, synchrotron, fluorescence, phosphorescence, explosion, biological, others
Geometry	Size, two- and three-dimensional beam shape, divergence, pointing, spatial coherence
Wavelength	X-ray, far and near ultraviolet, visible, near and far infrared
Spectrum	Bandwidth, distribution, temporal coherence, tuning
Intensity	Intrinsic luminance, brightness, spatial and temporal stability, noise
Power	Efficiency, electrical and thermal stabilization, cooling
Polarization	Type, orientation, degree
Stability	Frequency, pointing, intensity, beam shape
Optical components	
Lenses	Type, material, design, specification, compromises, speed, resolution, surface finish, tolerances, handling and mounting
Beamsplitters	Flatness, thickness, multiple reflection, parallelism
Filters	Bandwidth, tuning, edge roll-off, efficiency, reflections, absorption
Aberrations	Seidel, chromatic, higher order
Geometry	Complexity, aperture, stops, vignetting
Refraction	Refractive index, angular dependence of reflection, dispersion
Materials	Birefringence, transmissivity, homogeneity, strength, chemical stability
Coatings	Complexity, efficiency, materials, durability, scatter
Optical system	
Mechanical	Design, rigidity, robustness, beam control, adjustments (and ranges)
Alignment	Methods, sequence, sensitivity, criteria, stability, locking, interactions
Polarization	Necessity, preservation, type ('S', 'P', circular or elliptical)
Detectors	
Type	Array, line, single point, static (e.g. film), dynamic (e.g. electronic)
Principle	Photoelectric effect, permanent or reversible chemical change, thermal or mechanical effect
Mechanical	Sensitive area, packaging, thermal stabilization, protection
Detection	Capabilities, quantum efficiency, sensitivity, gain, noise
Speed	Dead time, read rate, fatigue, recovery time
Noise	Shot, dark, amplifier, statistical, after-pulsing

Table C5.17. Environmental considerations.

Mechanical	Vibration, shock, *g*-loads, rigidity, static and dynamic stresses, strength, materials, construction, pressure, extraneous damage, weight, size, cost
Thermal	Heat, cold, static/dynamic gradients, changes, distortion, misalignment, insulation, control
Radiation	Internal and external stray light, stops, filters, surface treatments
Aero-optics	Variable refractive distortions of wavefront
Noise	Electromagnetic or radio frequency interference (EMI, RFI), grounding, hardening, screening, isolation, spurious signals, power supplies

Table C5.18. Possible effects matrix.

Examples of properties	Aberrations, absorption, adjustments, ageing, alignment, blooming, breaking, calibration, changes, charging, chemistry, collimation, constancy, corrosion, cost, damage, design, detectivity, dimensions, dispersion, distortions, durability, efficiency, electromagnetic interference, errors, external stray light, field, flare, friction, fungal growth, ghosts, *g*-loads, ground loops, hard radiation, heat sources and sinks, insulation, internal stray light, isolation, Lagrange invariant, mounting methods, noise, nonlinear effects, optical feedback, outgassing, packaging, pointing, polarization, positioning, pressure, quality, quantum efficiency, quantum statistics, radio-frequency interference, reflections, refractions, rigidity, robustness, screening, sensitivity to everything, shock, signal and data formats, size, specifications, spectrum, stability, static and dynamic stresses, stiction, strength, temperature, testing, thermal gradients, transmission, vibration, vignetting, weight
Examples of components	Adhesives, apertures, baffles, beamsplitters, bearings, coatings, construction, detectors, electronics, Faraday and Kerr effect devices, fibres, filters, GRINs, health monitors, lenses, light sources, materials, mounts, locking, moving parts, polarizers, power supplies, software, stops, structures, surface treatments, thermal control, windows

Table C5.19. Design homilies.

- Software and power supplies give more trouble and consume more resources than can possibly be imagined; plan not to be disappointed by this
- Performance and ruggedization are better achieved by good initial design than by any number of later attempts at remediation
- Underspecified science leads to impossible engineering
- Communication, error checking, and conflict resolution are equally applicable to systems, hardware, software, scientists, engineers, managers, and customers who provide the funds
- Knowledge alone is not sufficient . . . choosing people with proper experience is essential
- Configuration control from the start, where design evolves most rapidly, may prevent errors; unconscious neglect makes later disasters almost inevitable
- A top-down straw-man design, starting at system level, and percolating down to isolate any area of uncertainty, is desirable as soon as possible
- Hardware and software must be modular, and should map isomorphically
- Optical, mechanical, interconnection, and functional layouts must be available early
- Everything is made of india-rubber—nothing is truly rigid
- The Internet is an extremely valuable source of ideas and information; however, since content is not typically peer reviewed, it must be independently corroborated
- Fifty hours in the lab will easily save one hour in the library
- Simplicity is a worthy goal, but oversimplification can invite problems
- Innovation is only acceptable where no possible method currently exists
- Quality is meeting the specifications
- If in doubt, make it stout, out of things you know about

Table C5.20. Attitude homilies.

- Talk to the customer
- Intend the excellence of your design
- Be honest, pragmatic, and courageous
- Study how others have approached and solved similar problems
- Test potential solutions and guidelines for relevance to *your* purposes
- Understand why something was done the way it was done, before committing to doing it differently
- Seek advice, comments, and recommendations; detractors are especially creative, and will be most helpful, if somewhat indirectly
- Consider all guidance carefully: suspect that which does not fit with intuition
- Never trust opinion, however informed, without supporting evidence
- If you are being brilliantly innovative, think about it harder, and insist others check it
- Question everything; analyse and understand the answers
- Check every assumption; check it again
- Historical reasons can vary from wise cautionary guidelines to unexamined prejudice
- Distinguish between intelligent compromises and short cuts; there are no short cuts
- Avoid moving parts; avoid nonstandard parts; if possible, avoid parts
- Buy the best and cry only once
- Hope is not a strategy

C5.10 Homilies from experience

Q. How do you succeed?	A. Good choices
Q. How do you make good choices?	A. Experience
Q. How do you get experience?	A. Bad choices

Many lessons may be encapsulated in aphorisms, whose illusion of triviality belies the pain associated with truly apprehending their significance. As light relief, two tables give examples of maxims encapsulating arts acquired by bitter experience. Table C5.19 passively addresses system design issues, whereas table C5.20 advances less scientifically tractable ideas and attitudes whose comprehension helps to reduce risk. These more psychologically based concepts are cautiously offered to mitigate the anguish incurred by those with the temerity to embark upon optoelectronic projects.

Further reading

The following bibliography merely skims a vast collection of interesting and useful works. The criterion for their choice is that each represents an in-depth study of one or more aspects of optoelectronics, which is either useful of itself or presents a method or analogy of more general application. While access to the Internet is now necessary, its content is not peer reviewed and may not be trusted without the corroborative evidence of which the bibliography includes examples.

Abbiss J B and Smart A E (ed) 1988 *Photon Correlation Techniques and Applications: OSA Conf. Proc. Series* vol 1 (Washington, DC: Optical Society of America)

Bass M (ed) 1995 *Handbook of Optics* vols 1–4 (New York: McGraw-Hill)

Berne B J and Pecora R 1976 *Dynamic Light Scattering* (New York: Wiley)

Born M and Wolf E 1980 *Principles of Optics* 6th edn (Oxford: Pergamon)

Brown N J 1986 Preparation of ultra-smooth surfaces *Annu. Rev. Mater. Sci.* **16** 371–388.

Brown R G W 1987 Dynamic light scattering using monomode optical fibers *Appl. Opt.* **26** 4846–4851

Brown R G W and Daniels M 1989 Characterization of silicon avalanche photodiodes for photon correlation measurements. 3: Sub-Geiger operation *Appl. Opt.* **28** 4616–4621

Brown R G W, Jones R, Ridley K D and Rarity J G 1987 Characterization of silicon avalanche photodiodes for photon correlation measurements. 2: Active quenching *Appl. Opt.* **26** 2383–2389

Brown R G W, Ridley K D and Rarity J G 1986 Characterization of silicon avalanche photodiodes for photon correlation measurements. 1: Passive quenching *Appl. Opt.* **25** 4122–4126

Brown W (ed) 1993 *Dynamic Light Scattering* (Oxford: Clarendon)

Chu B 1991 *Laser Light Scattering* 2nd edn (San Diego: Academic)

Dainty J C (ed) 1984 *Laser Speckle and Related Phenomena* (Berlin: Springer)

Danielsson L and Pike E R 1983 Long-range anemometry — a comparative review *J. Phys. E: Sci. Instrum.* **16** 107–118

Dautet H, Deschamps P, Dion B, MacGregor A D, MacSween D, McIntyre R J, Trottier C and Webb P P 1993 Photon counting techniques with silicon avalanche photodiodes *Appl. Opt.* **32** 3894–3900

Driscoll W G and Vaughan W (ed) 1978 *Handbook of Optics* (New York: McGraw-Hill)

Durst F Melling A and Whitelaw J H 1976 *Principles and Practice of Laser Doppler Anemometry* (London: Academic)

Hecht E 1987 *Optics* (Boston: Addison Wesley)

Johnson C S and Gabriel D A 1994 *Laser Light Scattering* (New York: Dover)

Kingslake R 1983 *Optical System Design* (New York: Academic)

Lasers, Photonics and Environmental Optics 20 August 2000 *Special Issue of Applied Optics* **40**

Lasers, Photonics and Environmental Optics 20 October 1997 *Special Issue of Applied Optics* **36**

Luxmoore A (ed) 1983 *Optical Transducers and Techniques in Engineering Measurement* (London: Applied Science)

Macleod H A 1986 *Thin Film Optical Filters* 2nd edn (London: Adam Hilger)

Mandel L and Wolf E 1995 *Optical Coherence and Quantum Optics* (Cambridge: Cambridge University Press)

Mayo W T and Smart A E (ed) 1980 *Photon Correlation Techniques in Fluid Mechanics* (Stanford: Stanford University)

Melles-Griot Catalog *The Practical Application of Light* current edn

MIL-HDBK-141 1962 *Military Standardization Handbook on Optical Design* (Washington DC: US Government Printing Office)

Minnaert M 1954 *Light and Color in the Open Air* (New York: Dover)

Photon Correlation and Scattering: Theory and Applications 20 July 1993 *Special Issue of Applied Optics* **32**

Saleh B E A and Teich M 1991 *Fundamentals of Photonics* (New York: Wiley–InterScience)

Schätzel K 1987 Correlation techniques in dynamic light scattering *Appl. Phys. B* **42** 193–213

Siegman A 1986 *Lasers* (Mill Valley, CA: University Science Books)

Smart A E 1991 Velocity sensor for an airborne optical air data system *AIAA: J. Aircraft* **28** 163–164

Smart A E 1992 Optical velocity sensor for air data applications *Opt. Eng.* **31** 166–173

Smart A E 1994 Folk wisdom in optical design *Optics and Photonics News, Engineering and Laboratory Notes 5*

Smith W 1972 *Modern Optical Engineering* (New York: McGraw-Hill)

Suparno, Deurbo K, Stamatelopolous P, Srivastra R and Thomas J C 1994 Light scattering with single mode fiber collimators *Appl. Opt.* **33** 7200–7205

van de Hulst H C 1981 *Light Scattering by Small Particles* (New York: Dover)

Vaughan J M 1989 *The Fabry–Perot Interferometer* (Bristol: Hilger)

Weast R C *Handbook of Chemistry and Physics* current edn (Boca Raton: CRC)

Welford W T and Winston R 1978 *The Optics of Nonimaging Concentrators* (New York: Academic)

Wyatt C L 1991 *Electro-Optical System Design for Information Processing* (New York: McGraw-Hill)

Yariv A 1995 *Optical Electronics* 5th edn (USA: Oxford University Press)

Author's Note: Many ideas and occasional short sections of modified text are reproduced from Brown R G W and Smart A E 1997 Practical considerations in photon correlation experiments *Appl. Opt.* **36** 7477–79, with kind permission of the Optical Society of America.